格致方法
定量研究系列

· 革新研究理念
· 丰富研究工具
· 最权威、最前沿的定量研究方法指南

吴晓刚 / 主编
【美】尤恩·苏尔李 (Eun Sul Lee) / 等著

数据分析方法五种

格致出版社 上海人民出版社

世纪出版

格致方法

定量研究系列

数据分析方法五种

本书由五种讨论数据分析方法的小册子组成，分别是《分析复杂调查数据》、《缺失数据》、《社会网络分析》、《因子分析：统计方法与应用问题》以及《基于行动者的模型》。本书主要介绍社会学研究方法之一，即数据分析方法。该书涵盖的是社会科学中技术性非常强的内容，前四种小册子从基于设计的分析入手，介绍了分析复杂调查数据的传统方法，阐述了如何处理社会科学研究中的缺失数据，同时对跨学科的社会网络分析法进行了说明，并详细介绍了因子分析法。最后，从基于模型的角度，以举例的形式详细解释了基于行动者的模型的基本原理，讨论了相关软件的使用，填补了该领域的空白。

上架建议：社科研究方法
ISBN 978-7-5432-1900-7
9 787543 219007 >
定价：75.00元
易文网：www.ewen.cc
格致网：www.hibooks.cn

格致方法
定量研究系列

吴晓刚 / 主编

【美】尤恩·苏尔李 (Eun Sul Lee) / 等著

数据分析方法五种

SAGE Publications ,Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

数据分析方法五种/(美)苏尔李(Lee, Eun Sul)等著;吴晓刚主编.—上海:格致出版社;上海人民出版社,2011
(格致方法·定量研究系列)
ISBN 978-7-5432-1900-7

Ⅰ.①数… Ⅱ.①苏… ②吴… Ⅲ.①统计数据-统计分析(数学) Ⅳ.①O212.1

中国版本图书馆CIP数据核字(2010)第254756号

责任编辑 王亚丽
封面装帧 人马艺术工作室·储平

格致方法·定量研究系列
数据分析方法五种
[美]尤恩·苏尔李 著
吴晓刚 主编

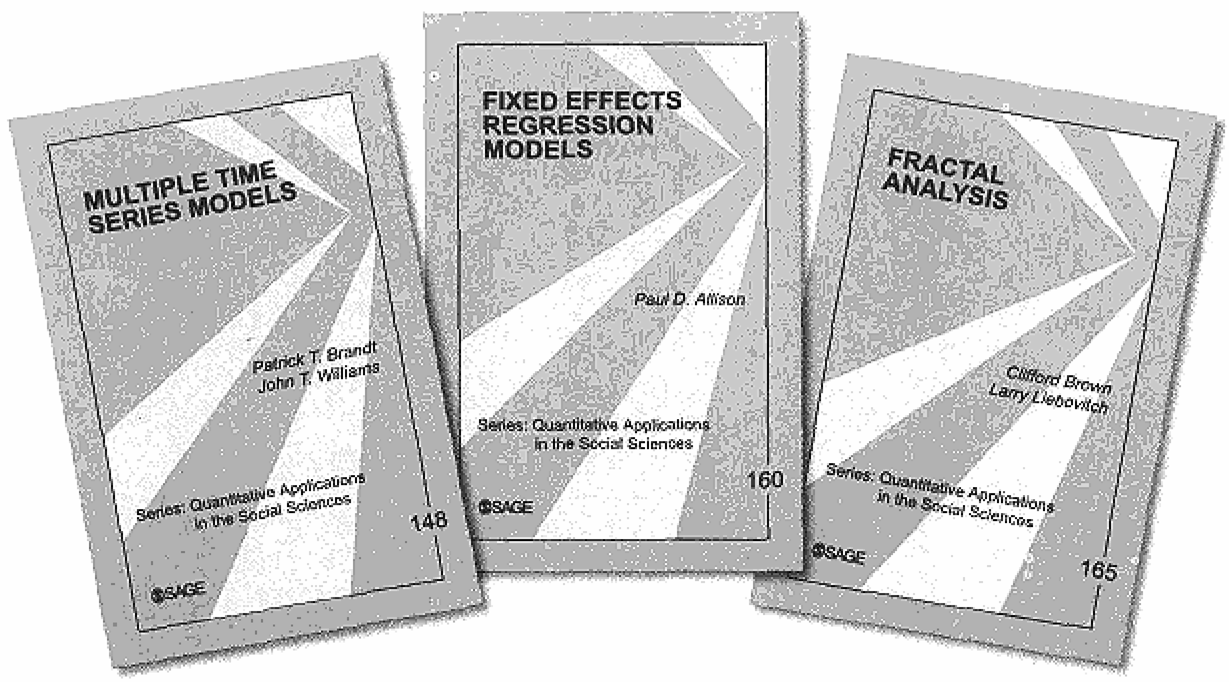
出　版　世纪出版集团　www.ewen.cc
格致出版社　www.hibooks.cn
上海人民出版社
(200001 上海福建中路193号24层)

编辑部热线 021-63914988
市场部热线 021-63914081

发　行　世纪出版集团发行中心
印　刷　上海图宇印刷有限公司
开　本　787×1092毫米 1/16
印　张　33.5
插　页　1
字　数　486,000
版　次　2011年9月第1版
印　次　2011年9月第1次印刷
ISBN 978-7-5432-1900-7/C·45
定　价　75.00元

SAGE

绿皮书的传奇



1976 年，当政府管理与政治学教授 Eric Uslaner 向 SAGE 出版社的创始人 Sara 建议出版关于重要定量研究方法的简明小册子时，没有人预见到这套书会取得巨大成功。

那年夏天，"社会科学定量研究方法" 丛书诞生。统一的、朴素的绿色封面，上面仅仅印着书名、作者名及本书所属的系列名。Iverson 和 Northporth 的《变量分析》(*Analysis of Variance*)、Nagel 的《运筹学》(*Operations Research*) 以及 Henkel 的《显著性检验》(*Tests of Significance*) 是最早出版的几本，每本售价 2.95 英镑，被形容为 "像烤饼一样好卖"。

SAGE 选择了 20 种顶尖的定量研究工具，然后去寻找合适的作者，邀请他们围绕这些工具撰写 92 页的小书。这些薄薄的绿皮书在课堂上深受欢迎，在图书馆成套陈列。

至今，绿皮书系列中共有 160 种在初版或再版。书的主题反映了量化研究方法的发展：从基础统计知识、数据类型、测量到计算机的应用以及博弈论。这套书非常畅销，其中最畅销的一本，是 1980 年出版的 Michael Lewis-Beck 的《回归方法的应用》(*Applied Regression*)。

出版说明

本书由五种讨论数据分析方法的小册子组成，分别是《分析复杂调查数据》、《缺失数据》、《社会网络分析》、《因子分析：统计方法与应用问题》以及《基于行动者的模型》。本书的主要内容如书名所示，是介绍社会学研究分析方法之一——数据分析方法。《分析复杂调查数据》为研究者在抽样设计和调查数据分析的联合方面提供了最新的情况，使研究者能深入了解方差估计的策略及抽样权重后复杂截面调查数据的初级和多变量分析等主要问题，使假设与研究相适应，从而完成更有效的统计推断。《缺失数据》给出了处理缺失数据问题的新策略，在缺失数据的条件下回顾最大似然估计，解释了插补法的EM算法和多重插补方法，并讨论了不可忽略的缺失数据。《社会网络分析》针对这种跨学科的数据分析的研究方法，阐述了社会网络分析的一些基本概念和数据收集问题，并介绍了网络分析方法和模型，讨论了社会网络分析的三个主要的潜在假设，反映了社会网络分析的最新进展。《因子分析：统计方法与应用问题》用明确的数据分析例子，详细地介绍了因子分析的不同方法，以及它们各自在何种情况下最实用。同时深入地讨论了验证性和探索性因子分析的差别，使读者能够对因子分析建立起系统性的知识。《基于行动者的模型》以举例的形式详细解释了基于行动者模型的基本原理，并讨论了基于行动者模型的应用软件，清晰明确，填补了该领域的空白。

数据分析方法虽然涵盖的是社会科学中技术性非常强的内容，但本书行文避免了复杂的数学公式推导。作者用具体的例子，结合语言和图形，帮助研究者理解相关分析方法的基本概念和模型。相信本书的出版能够推动国内社会科学定量研究的进一步发展。

总序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996 年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院（Institute for Social Research）定量社会研究方法项目的一些讨论会，受益良多。

2003年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。"社会科学里的统计学"(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。"社会科学中的定量分析"则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的"定量社会科学研究"丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选

一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

由于所选每本书都有一篇序言，对相关方法的背景和应用作了很好的介绍，我们均予以保留，内容在此不再赘述。为了方便起见，我们将内容相似的书目集册出版，每册三至五本不等，共八册，它们分别是：《线性回归分析基础》、

《高级回归分析》、《广义线性模型》、《列表数据分析》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》和《数据分析方法五种》。所冠书名未必能精准涵盖其中的内容,读者可自行参阅每本书的序言或目录。

我们希望本丛书的出版,能为推动国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

目 录

分析复杂调查数据

序 3
第 1 章　概论 5
第 2 章　抽样设计和调查数据 8
第 3 章　分析调查数据的复杂性 17
第 4 章　方差估计的策略 29
第 5 章　调查数据分析的准备 47
第 6 章　调查数据分析的操作 58
第 7 章　总结 88
注释 90
参考文献 92
译名对照表 98

缺失数据

序 105
第 1 章　导论 107

第 2 章　假设 109
第 3 章　传统的方法 112
第 4 章　最大似然 119
第 5 章　多重插补：基本原理 133
第 6 章　多重插补：复杂化 153
第 7 章　不可忽略的缺失数据 178
第 8 章　总结与结论 186
注释 188
参考文献 190
译名对照表 193

社会网络分析

序 199
第 1 章　社会网络分析简介 201
第 2 章　网络基础 204
第 3 章　数据收集 216

第 4 章　网络分析基本方法 248
第 5 章　网络分析高级方法 294
参考文献 320
译名对照表 331

因子分析：统计方法与应用问题

序 341
第 1 章　导论 343
第 2 章　抽取初始因子的方法 348
第 3 章　旋转的方法 364
第 4 章　再论因子数量的问题 376
第 5 章　验证性因子分析简介 381
第 6 章　建立因子尺度 395
第 7 章　对常见问题的简单回答 407
注释 414
参考文献 415

译名对照表 420

基于行动者的模型

序 425
前言 426
第 1 章　基于行动者建模的构想 428
第 2 章　行动者、环境和时间尺度 448
第 3 章　在社会科学研究中使用基于行动者的模型 458
第 4 章　设计和发展基于行动者的模型 475
第 5 章　基于行动者建模的发展 497
资源 505
注释 508
参考文献 509
译名对照表 520

分析复杂调查数据

作者简介

尤恩·苏尔李(Eun Sul Lee)

美国休斯敦得克萨斯大学健康科学中心公共卫生学院生物统计学教授，讲授健康调查的抽样技术和中级生物统计学课程。研究领域为抽样调查设计、有关健康的调查数据分析，以及在人口学和公共卫生领域的生命表应用和生存分析。与罗纳德 N. 福索佛(Ronald N. Forthofer)合著《生物统计简介：指南的设计、分析与探索》(*Introduction to Biostatistics: A Guide to Design, Analysis and Discovery*)一书(1995)。

罗纳德 N. 福索佛(Ronald N. Forthofer)

美国休斯敦得克萨斯大学健康科学中心公共卫生学院生物统计学退休教授。研究领域涉及线性模型的应用和健康研究中的类别数据分析技术等。与罗伯特·莱南(Robert Lehnen)合著《公共项目分析：一种新的分类数据方法》(*Public Program Analysis: A New Categorical Data Approach*)(1981)一书，与尤恩·苏尔李合著《生物统计简介：指南的设计、分析与探索》(*Introduction to Biostatistics: A Guide to Design, Analysis and Discovery*)(1995)一书。

关于译者

张卓妮

曾先后获得中国人民大学社会学法学学士学位和香港科技大学社会科学部硕士学位，现为香港科技大学社会科学部博士候选人。研究领域为社会分层与流动、中国的国内移民及跨国移民。

序

当乔治·盖洛普(George Gallup)正确预测到富兰克林·罗斯福(Franklin D. Roo-selvelt)成为1936年美国总统选举赢家时,公众民意调查进入了科学抽样的时代,那时使用的方法是配额抽样,一种代表目标总体的非概率抽样方法。但是,同样的方法却错误地预测托马斯·杜威(Thomas Dewey)将成为1948年的总统,但实际上哈利·杜鲁门(Harry S. Truman)才是最终的赢家。这种方法之所以失败,是因为配额抽样是非概率的,而且盖洛普的配额抽样框基于1940年的人口普查,忽略了二次大战期间的城市移民。

今天的调查抽样方法和早些时候相比已有了很大进步,现在我们依赖于复杂的概率抽样设计。一个关键的特征就是分层:目标总体被分成一系列不同阶层的子总体,阶层内的样本量由抽样者控制且通常与该阶层的人口规模成比例。另一个特征是整群和多阶段抽样:组群——作为在不同阶层选到的一系列不同等级的集群——先被选取,到最后那个阶层,个体成员才在这些"最后阶层"的集群内部被选中。比如,美国综合社会调查使用的就是分层多阶段集群抽样设计。Kalton(1983)为调查抽样提供了很好的入门介绍。

当调查设计具有这个复杂的性质时,数据的统计分析就不再是一个简单的运行回归(或任何其他的模型)分析的事情了。现在的调查全都伴随着抽样权重以支持正确的统计推断,大部分关于统计分析的文字,通过假定简单随机抽样而没有处理抽样权重的问题,而这个被忽略的抽样权重可能对统计推断产生重要的影响。在最近的二三十年间,数据分析的统计方法也已取得了巨大进展。这些很可能就是Michael S. lewis-Beck选择为《分析复杂调查数据》(*Analyzing Complex Survey Data*)出第二版的原因。

Lee和Forthofer这本书的第二版为我们在调查抽样设计和调查数据分析的联合方面提供了最新的情况。作者在本书开头先回顾了调查抽样设计常见的类型,并通过解释什么是抽样权重及其如何产生和调整,揭开了抽样权重的神秘面纱。然后他们详细讨论了方差估计以及考虑抽样权重后的复杂截面调查数据的初级和多变量分析等主要问题。

他们重点关注了基于设计的方法，这种方法直接在分析中涉及抽样设计（虽然他们也讨论了基于模型的视角，该视角在某些分析中能扩充基于设计的方法），他们还以流行的软件为例阐释了这种方法的使用。调查分析的学生将会在他们以样本为基础作出统计推断的努力中发现本书的巨大益处。

——廖福挺（Tim Futing Liao）

第 1 章 | 概论

调查分析似乎通常是在所有的样本观察值都以同等的概率被独立选中的方式操作实行的。如果在数据收集中使用的是简单随机抽样(SRS),这种分析就是对的,但实际上样本选择比 SRS 复杂得多。某些样本观察值可能比其他观察值以更高的概率被选中,之所以被包括在样本中,是因为他们属于某个特定组别的成员(比如家庭),而不是被独立选取的。我们可以在调查数据的分析中简单地忽略这些与 SRS 相悖的事实吗?使用调查数据分析统计书中的标准技术是否合适?或者是否有特别的方法和计算机程序可用于更合适的复杂调查数据的分析?这些问题将在随后的章节中进行论述。

今天典型的社会调查反映了统计理论和关于社会现象的知识两者间的结合,过去 70 年间从许多不同的调查中获得的经验塑造了它的发展。社会调查的进行是为了满足用以讨论社会、政治和公共卫生议题的信息需要。为了满足这种信息需要,在政府内部和外部都设立了调查机构。但是,在早期提供这些需要的信息的尝试中,调查小组们最关心的是实地调查中的操作问题,如抽样框的建立,员工培训/监督,以及成本的降低等,理论上的抽样议题获得的只是次级重视(Stephan, 1948)。随着这些实际问题得到解决,现代抽样实践方法的发展远远超越了 SRS。复杂抽样设计已经走在了前面,随之而来的是一系列分析问题。

因为早期调查通常需要的只是描述性统计,所以很少有人对分析问题感兴趣。更近一些时期,社会和政策科学家对分析研究的需要已经增加,那些并不参与数据收集过程的研究人员,正利用可用的社会调查数据对各种不同的现实议题进行分析。这个传统上被称为二手资料分析(Kendall & Lazarsfeld,

1950)。研究者通常并没有对复杂抽样设计的发展给予应有的关注,并假设这些设计对将要使用的分析程序没有什么影响。

二手资料分析中统计技术使用的增加以及近年来对数线性模型、logistic 回归和其他多变量技术的使用(Aldrich & Nelson, 1984; Goodman, 1972; Swafford, 1980),并没有把设计和分析更紧密地结合起来。这些技术断定数据收集使用的是简单有放回随机抽样(SRSWR),但这个假设在应用了观测单位的分层和整群方法以及不等概率选择法的社会调查中几乎无法得到满足。因此,利用 SRSWR 假设的社会调查分析可能导致有偏差和误导性的结果。比如,Kiecolt 和 Nathan(1985)在他们关于二手资料分析的书中承认了这个问题,但是他们几乎没有就如何把抽样权重和其他设计特征融入分析之中提供什么建议。最近一篇关于公共健康和流行病学的文献研究表明:对基于设计的调查分析方法的使用正在逐渐增加,但依然处于较低水平(Levy & Stolte, 2000)。

任何对抽样进行了限制且那些限制超越了 SRSWR 所做限制的调查在设计上都很复杂,并需要特别的分析性考虑。这本书回顾了由复杂抽样调查引起的分析问题,为分析策略提供了入门介绍,并利用某些可用的软件进行了实例阐释。我们讨论的重点在于使用抽样权重以校正不同的代表性,以及抽样设计对抽样方差估计的影响,我们也对权重的产生和调整程序做了一些讨论。但许多其他重要的处理非抽样误差和缺失数据的议题并没有在本书中得到充分说明。

本书介绍的最基本的方法是分析复杂调查数据的传统方法。这种方法现在被称为基于设计的(或基于随机化的)分析。另一种不同的分析复杂调查数据的方法,是所谓的基于模型的分析。正如统计学的其他领域,近些年来基于模型的统计推断在调查数据分析中已引起了更多的注意。这些构造模型的方法在调查数据分析的不同步骤——定义参数、定义估计值和估计方差——中得以介绍;但是,并没有普遍被接受的模型选择或使某个特定模型有效化的规则。

然而,对基于模型的方法的理解对调查数据分析者扩充基于设计的方法来说非常重要。在某些情况下,这两种方法会产生相同的结果,但在其他情况下结果却不同。基于模型的方法可能在描述性数据中没什么作用,但在推断性分析中却是有用的。我们将在适当的地方介绍基于模型的视角,并进一步提供与

这些问题的处理方式相关的参考资料。恰当地实施基于模型的分析需要对一般统计模型知识的掌握，而且还需要从调查统计员那里得到某些参考。此书中与这种不同的方法或相关题目有关的章节用星号(*)标出。

自本书第一版出版以来，对复杂调查数据进行分析的软件情况已有相当大的改进。方便用户的程序现在很容易就可获得，许多常用的统计方法现在也已合并成不同的程序包，包括 logistic 回归和生存分析。这些程序将在这个版本中用实例加以介绍。这些程序可能比其他标准软件更容易被错误地使用。本书所讨论的题目和议题将为避免调查数据分析中的陷阱提出一些指导原则。

在我们的叙述中，我们假定读者在一定程度上熟悉诸如简单随机抽样、系统抽样、分层随机抽样和简单二阶段整群抽样等抽样设计方法。对这些设计方法详细的描述可参见 Kalton(1983)和 Lohr(1999)的著作。我们还假定读者对标准统计方法以及某种标准统计程序包如 SAS 或 STATA 有大致的了解。

第 2 章 | 抽样设计和调查数据

我们对调查数据的考虑重点关注满足两个基本要求的样本设计。首先，我们只关心概率抽样，其中总体里面的每个成员均有已知的不等于 0 的概率被选中在样本里面。这是对某个给定的设计把统计理论应用到调查估计量的特性的推导中的基础。其次，如果某个样本要从某个总体中被抽取出来，就有必要建立一个罗列了合适抽样单位的抽样框，且这些抽样单位必须涵盖总体中的所有成员。如果罗列所有总体成员是不可行或不现实的，那么某些总体成员的集群也可用做抽样单位。比如，建立一个包括美国所有家庭的清单是不切实际的，但是我们可以在几个不同的阶段选择样本。第一个阶段，随机选取郡；第二个阶段，在每个被选中的郡内抽取普查区；第三个阶段，在每个被选中的普查区内选取街区。然后在最后的选择阶段，只需要对被选中的街区罗列家庭清单即可。这个多阶段设计保证了总体成员都有已知的且不等于 0 的被选中概率。

抽样方法的种类

最简单的抽样设计是简单随机抽样，它要求总体中的每个成员都有相同的被包括在样本里面的概率，且罗列了所有总体成员的清单是可以获得的。样本成员的选择可以通过有放回或无放回两种方式进行。简单有放回随机抽样特别让人感兴趣，因为它通过放回处理排除了被选中的成员间的相关性（协方差），因而简化了统计推断。但在这个方案中，一个成员可能在样本中出现不止一次。在实际中，简单随机抽样通过无放回的方式进行，即简单无放回随机抽样（SRSWOR），因为没有必要不止一次地从同一个成员中搜集信息。此外，

SRSWOR相比 SRSWR 给出的抽样方差更小。但是，对于只从总体成员中选取一小部分的大型调查来说，这两种方法实际上是一样的。我们将在本书中使用 SRS 一词表示 SRSWOR，除非另有说明。

SRS 设计被进一步调整以适应其他理论或实际的考虑。常用的实际设计方案包括系统抽样、分层随机抽样、多阶段集群抽样、PPS 抽样（按规模大小成比例概率抽样）和其他受控制的选择程序。这些更实际的设计在两个重要方面偏离了 SRS。首先，总体成员被选中的概率（以及由个体成员组成的不同集群的联合被选中概率）可能并不相同。其次，抽样单位可能不同于我们感兴趣的总体成员。这些偏离使常用的估计方法和方差计算方法复杂化，而且如果没有使用恰当的分析方法，可能会导致估计和统计检验的偏差。我们将用几个具体的抽样设计详细考虑这些偏离情况，并检验它们对调查分析的影响。

系统抽样是替代 SRS 的一种比较常用的方法，其原因在于它的简单性。它在一个随机起始点（在 1 和 k 之间）之后，选择每个处于第 k 个位置的成员。其程序上的任务很简单，而且我们很容易对该程序进行检查，但是要通过检验结果来核实 SRS 却很困难。它常用于多阶段抽样的最后一个阶段，如当实际调查者被指示要从某个街区的住宅清单中选择预定比例的（住宅）单位时。系统抽样程序对总体中的每个成员指定相同的被选中的概率，这确保当总体中的成员数量 N 等于样本中成员数量 n 的 k 倍时，样本均值将是总体均值的无偏估计。如果 N 并不严格等于 nk，就无法确保相同的概率，虽然这个问题在 N 很大时可以被忽略。在这种情况下，我们可以使用循环系统抽样方案。在这个方案中，随机起始点是在 1 至 N 之间选择的（任何成员都可以成为起始点），然后假定这个抽样框是循环的（列表的末端与列表的开头是相连的），选择每个处于第 k 个位置的成员。但是，当抽样框中的成员是以与调查变量相关的循环方式罗列，且抽样间距正好与清单的循环周期相吻合时，系统抽样就可能导致不真实的估计。比如，如果我们对到某个诊所的每隔 40 个病人中抽取 1 个，而平均每天的病人流量几乎就是 40，那么得到的系统抽样将只包括那些在一天中的某个特定时间到该诊所的人。这样的样本可能无法代表这个诊所的所有病人。

再者，即使当清单是随机排序时，与 SRS 不同，总体成员的不同的集合也可

能具有不同的被选中的概率。比如，在系统抽样中，同时包含第 i 个和第 $i+k$ 个成员在样本中的概率是 $1/k$，但同时包含第 i 个和 $i+k+1$ 个成员在样本中的概率为 0。这使方差计算复杂化了。另一个观察系统抽样的方式是，相当于从 k 个有规则地形成的包含 n 个成员的集群中选取一个。(不同集群之间的)抽样方差无法从选到的这个集群中估计出来。因此，系统抽样中的方差估计需要特殊的策略。

用于克服这些系统抽样问题的一种修正方法是重复系统抽样(Levy & Lemeshow, 1999:101—110)。这种方法不是只从清单中进行一次抽样来选取一个系统的样本，而是分几次选取几个更小的系统性的样本，而每一次选取都以不同的起始点从头到尾进行选择。这个程序不仅避免了可能出现的抽样框的周期性问题，而且允许直接从数据中进行方差估计。从所有子样本中得到的估计值的方差可以通过从每个子样本中各自的不同估计值的变异情况中估计出来。重复系统抽样的方法为复杂调查的方差估计提供了一种策略，这将在第 4 章进一步讨论。

分层随机抽样把总体成员划成不同的阶层，每个阶层里又有不同的样本。它的使用有几个原因：(1)如果阶层是内部同质的，则该方法可以减少抽样方差；(2)对不同的阶层可以获得不同的估计值；(3)利用阶层，实地调查的执行可以组织起来；(4)在不同的阶层内部，不同的抽样需要可以得到满足。

当不同阶层间抽样的比例是统一的时候，样本在不同阶层间的分配是成比例的；或者，当某个更高的抽样比例应用到某个更小的阶层，从而为比较研究选取足够的对象时，样本分配是不成比例的。总的来说，分层随机抽样的估计过程比 SRS 更加复杂。通常被称为两步处理法。第一步是分别在每个阶层内部进行统计量——比如均值及其方差——的计算。然后以反映每个阶层内总体比例的权重为基础把这些估计值结合起来。正如我们后面将要讨论的，这种方法也可称为利用加权统计量的一步处理法。在成比例分层抽样的情况下，估计过程可以简化，但在方差估计时必须把阶层考虑进来。

划分阶层需要在抽样框中找到用来分层的变量的信息。当这些信息找不到时，分层方法就不能应用到抽样设计中。但可以在数据搜集完成之后进行分

层以改善估计的精确度。这种事后分层方法的使用，通过把样本的人口构成调整到可知的总体构成水平，从而使样本更具总体代表性。通常，为了利用人口普查数据，在事后分层中常使用诸如年龄、性别、种族和教育等人口统计学的变量。这种调整需要使用权重和不同的方差估计方法，因为在事后分层设计中，阶层的样本规模是一个随机变量(在数据收集完之后才决定的)。

集群抽样是一种实用的调查方法，因为它以由不同的成员构成的组(集群)为单位，而不是直接对个体成员进行抽样。它简化了构造抽样框的任务，降低了调查成本。正如前面所讨论的，通常使用的是一系列不同阶层的地理集群。在多阶段集群抽样中，除了在最后的抽样阶段，其他阶段中的抽样单位是由成员构成的组别。当成员的数量在不同的集群间相同的时候，估计过程也与 SRS 相同。但集群规模不同的简单随机抽样会导致处于更小集群中的成员比处于更大集群中的成员更有可能被选中。此外，集群通常被分层以实现某些调查目的和实地的调查步骤，比如，对特殊少数人口集群的过取样。不成比例分层和不同规模集群使估计过程变得复杂。

在不同规模集群的一步集群抽样中选取自我加权的成员样本，其中一个方法是以与集群规模成比例的概率选取集群(PPS 抽样)。但是，这需要知道集群的真实规模。因为真实规模通常在调查时并不可知，所以选择概率反而是与估计到的规模成比例的(PPES 抽样)。比如，在某个以医院作为集群的出院调查中，病床的数量可以用来测量医院规模。PPES 抽样的一个重要的后果是预期样本量将因初级抽样单位(PSU)的不同而不同。也就是说，样本量不是固定的，而是随着样本的改变而改变的。所以，样本量——样本均值计算中的分母——是一个随机变量，因此样本均值成为两个随机变量的比率。这种类型的变量，比率变量，需要特别的策略进行方差估计。

调查数据的属性

如果我们要从样本推断总体，样本选择过程就是推断过程中一个不可或缺的部分，且调查数据必须包含选择程序各重要层面的信息。考虑到多数社会调

查对 SRS 的偏离情况，我们不仅要把调查数据看成是测量的记录，还要注意它具有不同的代表性和结构安排。

抽样权重用于反映样本成员不同的被选择概率。抽样权重的产生需要密切关注每个抽样阶段和每个阶层各自的被选择概率。此外，它还涉及在样本的组别内校正不同的应答率，并通过人口统计学的变量把样本分布调整到已知的总体分布水平，即事后分层调整。还有，不同的分析单位可能需要不同的抽样权重。比如，在某个社区调查中，可能需要对个人层次的分析创建个人权重，对家庭数据的分析创建家庭权重。

当以下某种自加权设计被使用时，我们也可能会对未加入权重比较放心。一步集群抽样中正确的 PPS 抽样会产生一个自加权成员样本，就像在 SRS 设计里面一样。自加权也可以在两步设计中实现，即当在第一阶段中使用了正确的 PPS 抽样，且在每个被选中的 PSU 中选取固定数量的成员时。如果简单随机抽样在第一阶段中使用，且在第二阶段选取固定比例的成员，也会出现相同的结果(Kalton，1983，第 5 章和第 6 章)。但在实际中，自加权特征会遭到无应答和抽样框中的可能误差的破坏。这种非计划中的自加权可能产生偏差，但这种偏差却不大可能通过对样本数据进行检验来评估。事后分层和无应答调整是用于减少这种误差的两种方法。事后分层涉及权重的分配，以把不同子群中的样本比例调整为与相同子群的总体比例相一致。无应答调整对那些参与了调查的人提高了权重，以弥补那些具相似特征但没有应答的人。因为要通过加权而实现无应答和事后分层调整，所以即使当自加权设计被应用时权重的使用仍几乎不可避免。

抽样设计影响了标准误差的估计，因此也必须结合到分析中。仔细检查那些我们熟悉的、在统计教科书中出现且内置于大部分计算机程序包中的标准误差计算公式，我们就能发现它们都是基于 SRSWR 设计的。这些公式相对比较简单，因为在假设了成员的选择具独立性之后，成员间的协方差为 0。但对这些公式如何加以修正以调整其他复杂的抽样设计却并不清楚。

为了更好地理解对方差公式进行调整的需要，我们对几个不同的抽样设计检验一下方差公式。首先考虑从 SRSWOR 设计中得到的样本均值的方差。那

个我们熟悉的样本均值 $\bar{y}$ 的方差公式(通过 SRSWR 从有 N 个成员、均值为 $\overline{Y}$ 的总体中抽取一个包括 n 个成员的样本)在初级统计教科书中等于 σ^2/n,其中 $\sigma^2=\sum(Y_i-\overline{Y})^2/N$。这个公式在 SRSWOR 设计中需要被修改,因为这里一个成员的选择已经不再独立于另一个成员的选择。由于条件不允许重复选择,在第 i 个和第 j 个样本成员间存在一个负的协方差 $[-\sigma^2/(N-1)]$。合并 $n(n-1)$ 乘以协方差,SRSWOR 样本均值的方差就变成了 $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$,这个比 SRSWR 中的方差小,是其 $(N-n)/(N-1)$ 倍。把 σ^2 的无偏估计值用 $[(N-1)s^2/N]$ 代替,SRSWOR 样本均值的方差估计值就变成

$$\hat{V}(\bar{y})=\frac{s^2}{n}(1-f) \qquad [2.1]$$

其中 $s^2=\sum(x_i-\bar{x})^2/(n-1)$,$f=n/N$。$(N-n)/(N-1)$ 和 $(1-f)$ 两者都称为有限总体校正(FPC)因素。在一个大总体中,协方差会非常小,因为抽样部分非常小。因此 SRSWR 和 SRSWOR 设计会产生几乎相同的方差,这两种程序在所有实际目标中可以被看做是相同的。

分层抽样通常被认为是一个更有效的设计,因为如果使用恰当,它给出的方差比类似的 SRS 给出的方差更小。因为阶层间的协方差为 0,所以样本估计值的方差是从阶层内的变异情况——以阶层的样本量和阶层权重为基础结合起来——推导出来的。分层样本方差的取值取决于阶层样本量的分布。一个最优(或者 Neyman)分配产生的抽样方差要小于或等于基于 SRS 而产生的方差,但在极端少数的情况下也有例外。对其他不成比例的分配来说,当阶层内的有限总体校正因素不能被忽视时,抽样方差可能比那些基于 SRS 的方差要大。因此可以认定,相比 SRS,分层的方法总会降低抽样方差。

集群抽样设计常常导致比从 SRS 中得到的方差更大的抽样方差。这是因为在自然形成的集群内部其成员通常比较相似,因此在集群内各成员间产生一个正的协相关。集群内的同质性用组内相关系数(ICC)——集群内所有可能的成员配对组合间的相关性——测量。如果集群是随机形成的(即如果每个集群

都是成员的随机样本)，ICC 就等于 0。在许多自然集群中，ICC 是正的，因此抽样方差会比从 SRS 设计中得到的更大。

在一个复杂的设计中，很难就抽样方差的相对大小进行归纳，因为必需对分层和集群的联合效应以及抽样权重的效应进行评估。所以，调查数据中所有的观察值都必须看做是一个特定的包含抽样权重和结构安排的抽样设计的产物。除了抽样权重以外，阶层和集群的标识(至少 PSU)也应该包含在样本调查数据中。这些要求的原因我们也会在下文中详细论述。

复杂调查中方差计算的一个复杂情况来源于权重的使用。因为任何加权后方差估计之分母中的权重总和并不固定，反而随样本的不同而改变，所以该估计就变成了两个随机变量的比率。一般来说，比率估计值是有偏差的，但如果权重的变化相对较小或样本量比较大，这个偏差就可以忽略不计(Cochran, 1977，第 6 章)。因此，在大型社会调查中，比率估计的有偏差问题并不严重。但是，因为这个偏差，使用均方差——方差和偏差平方的和——比使用方差更合适。但因为这个偏差通常是可忽略的，因此在本书中我们使用"方差"一词，尽管其所指实为均方差。

调查数据的另一种不同看法*

迄今为止，对调查数据属性的描述是从基于设计的视角展开的——即样本数据是利用某个特定的样本选择设计方案从一个有限的总体中抽取的观察值。抽样设计明确了每个潜在样本的选择概率，然后某个合适的估计方法被选中用以反映这种设计。如概论中所述，基于模型的视角提供了样本调查数据的另一种看法。有限总体中的观察值被看做是从某个模型中产生的随机变量(服从某种概率分布的随机变量)的现实对应值。这假定的概率模型连接了样本中的单位和不在样本中的单位。在基于模型的方法中，样本数据用于预测无法观测到的值，因此推论可以被认为是预测问题(Royall, 1970, 1973)。

这两种不同的视角在可以合理地假设样本观测值是在一个均值为 μ、方差为 σ 的正态分布中独立同分布的 SRS 中可能没什么差别，但从模型的角度看，全及总

体是样本中的观测值的和以及没有包含在样本中的观测值的和，也就是 $Y = \sum_{i \in S} y_i + \sum_{i \notin S} y_i$。基于同均值的假设，全及总体的估计值可以写成 $\hat{Y} = n\bar{y} + (N-n)\bar{y} = N\bar{y}$，其中 $\bar{y}$ 是在该模型下未观测到的观察值的最优无偏估计值。结果它与基于设计的方法中的扩展估计值相同，即 $\hat{Y} = (N/n)\sum_{i=1}^{n} y_i = N\bar{y}$，其中$(N/n)$是抽样权重(SRS 中选择概率的倒数)。这两种方法都得出了同样的方差估计值(Lohr, 1999, 第 2.8 节)。

但是，如果采用的是不同的模型，方差估计值就可能不同。比如，在 SRS 比率[1]和回归估计的情况下，假定的模型是 $Y_i = \beta x_i + \varepsilon_i$，其中 Y_i 代表一个随机变量，x_i 是一个辅助变量，对这个辅助变量来说全及总体是已知的。在这个模型下，全及总体的线性估计为 $\hat{Y} = \sum_{i \in S} y_i + \sum_{i \notin S} y_i = n\bar{y} + \hat{\beta}\sum_{i \notin S} x_i$。其中第一部分是从样本中来的，第二部分是基于假定的模型对没有观察到的单位作出的估计。如果我们把 $\hat{\beta}$ 取值为样本比率 $\bar{y}/\bar{x}$，那么就有 $\hat{Y} = n\bar{y} + \frac{\bar{y}}{\bar{x}}\sum_{i \notin S} x_i = \frac{\bar{y}}{\bar{x}}(n\bar{x} + \sum_{i \notin S} x_i) = \frac{\bar{y}}{\bar{x}}X$，其中 X 是 x_i 的全及总体。这就是 Y 的比率估计。如果我们把 $\hat{\beta}$ 取值为估计到的回归系数，那么我们就得到了一个回归估计。虽然从基于设计的角度看比率估计被认为是轻微有偏的，但从基于模型的推理角度看如果模型正确它就是无偏的。

但是，从基于模型的方法中得到的方差估计值与从基于设计的方法中得到的稍微有所不同。基于设计的预计全及总体的方差估计值为 $\hat{V}_D(\hat{Y}) = \left(1-\frac{n}{N}\right)\left(\frac{N^2}{n}\right)\frac{\sum[y_i - (\bar{y}/\bar{x})x_i]^2}{n-1}$。基于模型的估计值为 $\hat{V}_M(\hat{Y}) = \left(1-\frac{x}{X}\right)\left(\frac{X^2}{x}\right)\frac{\sum[\{y_i - (\bar{y}/\bar{x})\}/\sqrt{x_i}]^2}{n-1}$，其中 x 是全及样本，X 是辅助变量的全及总体(Lohr, 1999, 第 3.4 节)。

当(a)y_i 和 x_i 之间的关系是过原点直线的，(b)y_i 关于这条线的方差是与 x_i 成比例时，比率估计模型是有效的。当 y_i 和 x_i 之间的相关性小于 x_i 的变异系数与 y_i 变异系数的比率的一半时，比率估计被认为差于(没有辅助变量的)扩

展估计(Cochran, 1977,第6章)。因此,在调查分析中使用比率估计要求对模型假设进行检验。在实际中,当数据库包括大批变量时,比率估计会变得笨重,要为不同的估计选择不同的辅助变量。

要把基于模型的方法应用到实际问题中,我们必须能够建立一个适当的模型。如果模型是错的,那么基于模型的估计就是有偏的。在抽样中使用基于模型的推断时,我们需要通过详细分析数据来检验模型的假设。在许多情况下,检验假设是困难的。模型的恰当性在某种程度上说是判断力的问题,而且适合某个分析的模型,可能对另一个分析或另一个调查并不合适。

第 3 章 | 分析调查数据的复杂性

调查数据分析的两个根本方面是调整样本观察值的不同的代表性和评估由样本设计选择的复杂性带来的精确度的得失。这一章介绍权重的概念并讨论样本设计的选择对方差估计的影响。为了阐释权重在调查分析中的多用性，我们列举了两个创建和调整抽样权重的实例。

调整不同的代表性：权重

有两种类型的权重在调查数据的分析中经常遇到：(1)扩展权重，它是选择概率的倒数；(2)相对权重，通过按比例缩小扩展权重以反映样本规模而获得。这一部分我们将为几种不同的抽样设计详细介绍这两种不同的权重。

考虑如下 SRS 情况：对一个清单中某总体的 $N=4000$ 个成员用 1 到 4000 进行排号。用一张随机数字表从这个总体中抽取固定数量的成员（比如 $n=200$），不允许重复选择（无放回）。选择概率或抽样比例为 $f=n/N=0.5$。扩展权重是选择概率的倒数，$w_i=1/f=N/n=20(i=1, \cdots, n)$，表示由一个样本观察值所代表的总体中的成员数量。这 n 个被选中的成员的权重之和等于 N。基于样本成员，变量 Y 的全及总体的估计值为

$$\hat{Y}=\sum w_i y_i=(N/n)\sum y_i=N\bar{y} \qquad [3.1]$$

方程 3.1 表现了扩展权重在样本观察值的加权总和中的使用。由于对 SRS 中的每个成员权重是一样的，因此这个估计值可以简化为样本均值的 N 倍（方程 3.1 中的最后一项）。同样，总体均值的估计值等于 $\hat{\bar{Y}}=\sum w_i y_i/\sum w_i$，即加

权后的样本均值。在 SRS 中，它可简化为 $(N/n)\sum y_i/N=\bar{y}$，表示样本均值是总体均值的估计值。但是，即使在不等概率设计中，各成员的权重也不相同，估计值仍然是全及总体的加权总和和总体均值的加权平均数。

虽然扩展权重看起来适合全及总体的估计，但它可能会给样本均值和其他统计测量带来重大影响。比如，在列联表中用扩展权重的总和代替以样本规模为基础的相对频数可能会夸大数据中的置信度。要处理这个问题，可以把扩展权重按比例缩小以生成相对权重 $(rw)_i$，即用扩展权重除以扩展权重的均值，$w_i/\bar{w}$。其中，$\bar{w}=\sum w_i/n$。样本中每个成员的相对权重之和等于 n。对 SRS 设计来说，每个成员的 $(rw)_i=1$。用相对权重进行加权后的全及总体的估计为

$$\hat{Y}=\bar{w}\sum(rw)_i y_i=(N/n)\sum y_i=N\bar{y} \qquad [3.2]$$

注意在方程 3.2 中，相对权重被乘以扩展权重的平均数，对 SRS 得出的简化估计值与方程 3.1 一模一样。因此，在对全及总体进行估计时，扩展权重的使用比相对权重更简单。相对权重在分析性研究中更合适，却不适用于估计总体和计算有限总体校正值。

大部分政府部门和调查机构提供的公开调查数据使用的都是扩展权重，很容易转化为相对权重。但这种转化并不是必须的，因为许多便于用户使用的调查分析统计程序会在适当的时候自动对其进行转化。

我们来分析一下分层随机抽样设计中扩展权重的使用。在这种设计中，具有 N 个成员的总体基于某个变量被分成 L 个阶层，每个阶层分别包含 N_1，N_2，…，N_L 个成员。从这些阶层中，分别有 n_h 个($h=1, 2, \cdots, L$)成员从第 h 个阶层中被抽取出来。当每个阶层的抽样比例相同的时候，这个分层抽样设计便具有自加权的特征。如果一共有 200 个成员从两个阶层 ($N_1=600$ 和 $N_2=3400$) 中按比例被抽取出来，那么抽样比例 $f=200/4000=0.05$。按这个比例选取(从每个阶层中各选取 5%的样本)的结果会产生两个样本 $n_1=30$ 和 $n_2=170$，因为 $f_1=30/600=0.5$，$f_2=170/3400=0.5$。因此加权的方案与 SRS 设计中的结果完全相同。

在不成比例分层抽样设计中情况则稍微不同。比如，如果这200个样本成员在两个阶层中平摊，那么 f_1（=100/600）和 f_2（=100/3400）的取值就不相同，这两个阶层的成员其扩展权重也不相同，分别为 $w_{1i}=6$ 和 $w_{2i}=34$。扩展权重之和在第一个阶层内为600，在第二个阶层内为3400，总和4000等于总体规模。扩展权重的均值 $\bar{w}=(100\times6+100\times34)/200=20$，因此相对权重之和在第一个阶层内为30，在第二个阶层内为170，两个阶层总和等于样本规模。值得注意的是，这两种权重的使用都相当于用总体在各阶层的分布状况对阶层均值（$\bar{y}_h$）进行加权（即 $\sum(N_h/N)\bar{y}_h$，标准程序）。在阶层1内，扩展权重和相对权重两者的总和都分别等于它们各自的全部总和的15%，而且第一阶层包含的成员的数量也等于所有总体成员数量的15%。

虽然我们在介绍抽样权重时用的是SRS和分层抽样设计，但同样的概念可以很容易扩展到更加复杂的设计的使用中。总的来说，抽样权重是选择概率的倒数，虽然我们经常会通过事后分层和无应答调整的方法对它做进一步的调整。对每个样本成员赋予抽样权重有助于对所有的抽样设计建立一个综合的估计程序。一个普遍的规则是，在调查数据分析中所有的估计都以加了权的统计量的形式进行。除了估计总体和计算有限总体校正值，这些权重的规模在估计参数和标准误差时并不重要。

接下来，我们举两个创建/调整抽样权重的例子。第一个例子告诉我们如何通过事后分层的方法调整抽样权重，以使样本结构与总体结构一致。同样的方法也可用于调整各不同人口子群中的不同的应答率。当涉及的人口特征变量只有少数几个的时候，事后分层的方法效果较好。第二个例子告诉我们在追踪调查中，我们可以通过使用logistic回归模型，对许多变量的不同的损耗率进行调整。

用事后分层的方法加权

为了演示抽样权重的创建过程，我们利用美国1984年的综合社会调查

(GSS)进行说明。这个调查是由美国全国民意研究中心(NORC)组织的一个复杂的抽样调查,旨在从美国普通的非居住于集体机构的成年(大于或等于18岁)人口中获得综合社会信息。该调查使用了多阶段抽样方案以在家庭层面产生一个自加权样本。然后从每一个被抽中的家庭里面随机选取一个成年人(Davis & Smith, 1985)。最终的数据总共包括1473个观察记录可供分析使用。对家庭层面的数据来说,扩展权重可以计算为等于美国所有家庭的总数量除以1473,而被选中的家庭内的扩展权重则等于该家庭内成年人的数量。这两种权重的乘积等于样本中个体成员的扩展权重。

为了分析GSS数据,我们只需要集中讨论家庭内部的权重,因为每个家庭都具有相同的被选中的概率。个体成员的相对概率可以计算为等于这个家庭内成年人的数量除以平均每个家庭内成年人的数量(后者为2852/1473 = 1.94)。这个权重反映了在维持样本规模的基础上,一个成年人被选取到样本中的概率。

为了使样本结构与总体结构一致,我们需进一步对这个权重进行调整。这样可以提高估计的准确度,同时也可把无应答和样本选择偏差降低到只与人口构成相关的程度。[2]如表3.1所示,调整因素的产生是为了使样本中个体的分布状况在年龄、种族和性别等方面符合1984年美国人口总体的分布情况。第1列显示美国人口普查局估计的分年龄、种族和性别的人口分布状况。第2列显示在对应的人口分类中,在选中的所有家庭里面加权后的成人数量,第3列显示的是对应的分布比例。调整因素是第1列和第3列数字的比。调整后的权重等于把调整因素乘以相对权重,这样调整后的权重的分布就与人口总体的分布一致。

这些调整因素表明,如果没有进行调整,那么GSS样本中25—34岁男性的比例低于实际比例,而非白种45—54岁男性和非白种25—34岁女性的样本比例则高于实际比例。

然后将这些调整后的相对权重用于数据分析中的成年人的比例。比如,用于肯定回答这样一个问题:"是否有你能够想象到的这样一类情况,在这类情况下你会同意一位男性攻击另一位陌生男性?"正如表3.2上半部显示的,加权后

表 3.1 事后分层调整因素的产生：美国 1984 年综合社会调查

统计群	人口分布 (1)	成人权重数 (2)	样本分布 (3)	调整因素 (4)
白人，男性				
18—24 岁	0.0719660	211	0.0739832	0.9727346
25—34	0.1028236	193	0.0676718	1.5194460
35—44	0.0708987	277	0.0795933	0.8907624
45—54	0.0557924	135	0.0473352	1.1786660
55—64	0.0544026	144	0.0504909	1.0774730
65 岁及以上	0.0574872	138	0.0483871	1.1880687
白人，女性				
18—24 岁	0.0705058	198	0.0694250	1.0155668
25—34	0.1007594	324	0.1136045	0.8869317
35—44	0.0777364	267	0.0936185	0.8303528
45—54	0.0582026	196	0.0682737	0.8469074
55—64	0.0610057	186	0.0652174	0.9354210
65 岁及以上	0.0823047	216	0.0757363	1.0867272
非白人，男性				
18—24 岁	0.0138044	34	0.0119215	1.1579480
25—34	0.0172057	30	0.0105189	1.6356880
35—44	0.0109779	30	0.0105189	1.0436290
45—54	0.0077643	37	0.0129734	0.5984774
55—64	0.0064683	12	0.0042076	1.5372900
65 岁及以上	0.0062688	18	0.0063113	0.9932661
非白人，女性				
18—24 岁	0.0145081	42	0.0145081	0.9851716
25—34	0.0196276	86	0.0301543	0.6509067
35—44	0.0130655	38	0.0133240	0.9806026
45—54	0.0094590	33	0.0115708	0.8174890
55—64	0.0079636	30	0.0105189	0.7570769
65 岁及以上	0.0090016	27	0.0094670	0.9508398
总计	1.0000000	2852	1.0000000	

资料来源：U. S. Bureau of the Census, Estimates of the population of the United States, by age, sex, and race, 1980 to 1985(Current Population Reports, Series P-25, No. 985), April 1986. Noninstitutional population estimates are derived from the estimated total population of 1984 (Table 1), adjusted by applying the ratio of noninstitutional to total population(Table A1).

的总体比例为 60.0%，比未加权时的估计值 59.4%稍微大一点。对人口子群体的估计结果显示，加权后估计值和未加权估计值之间的差异很小。这可能

主要归因于自加权的特征——反映在大部分家庭只有两个成年人这个现实上，以及在更小程度上反映在“同意攻击”与家庭内成年人数量并没有关系这个事实上。在美国国家精神卫生研究所发起的流行病学调查(Epidemiologic Catchment Area，ECA)中，情况则有所不同。如表 3.2 所示，在这个调查中，加权后的所有障碍和焦虑性障碍流行情况的估计值分别比未加权估计值低 20%和 26%。

表 3.2　两个调查中加权后估计和未加权估计的比较

调查及变量名称	加权后估计	未加权估计
1. 美国综合社会调查(GSS)		
(同意“攻击”的比例)		
总体	60.0	59.4
分性别		
男性	63.5	63.2
女性	56.8	56.8
分教育程度		
大学	68.7	68.6
高中	63.3	63.2
其他	46.8	45.2
2. 美国国家精神卫生研究所流行病学责任区调查(Epidemiologic Catchment Areas Survey)(精神障碍流行率)		
所有障碍	14.8	18.5
焦虑性障碍	6.5	8.8①

资料来源：Epidemiologic Catchment Areas Survey，数据来自 Lee，Forthofer & Lorimor(1986)，表 1。

最后，应该对这些调整后的权重进行检查，看看是否存在极端大的取值。极端的离异值可能意味着某些事后分层的样本规模太小以至于不可靠。在这种情况下，某些小的事后分层可能需要合并起来，或者需要使用某些排列程序以消除那些极端值(Little & Rubin，1987：59—60)。

① 原书中未加权估计的数据，所有障碍为 8.8，焦虑性障碍为 18.5，原书有误。——译者注

在追踪调查中调整权重

追踪调查在社会科学研究中很常用。但遗憾的是，对所有的初始受访者进行持续的跟踪几乎不可能。他们中的某些人可能已经去世、搬迁，或者拒绝参与追踪调查。这些事件并不是随机分布的，在追踪调查中不同的损耗率可能导致选择偏差。基于一些人口特征变量，我们可以像在事后分层中那样采用相同的方式进行调整，但我们可充分利用初始调查中大量的潜在估计变量进行追踪调查中的损耗调整。分层的策略可能并不是很适合大量的估计变量，但是 logistic 回归模型提供了一种方式以包含若干变量。以初始调查数据为基础，我们可以建立一个 logistic 回归模型，通过使用初始调查中的一系列精心挑选过的预测变量和常用的人口特征变量来预测损耗（二分类变量）。然后，基于原始样本中受访者特征的 logit 估计值可以用于调整在追踪调查中每个受访者的初始权重。实际上，这种方法可以通过不同程度地调高那些在追踪调查中成功接触到的人的权重，来弥补那些丢失掉的受访者，以此排除选择偏差，以使之降低到只与模型中的变量相关的程度。随着模型使用更多恰当的变量，损耗调整可以比初始调查中的无应答调整更加有效。对由“追踪调查中的丢失”带来的无应答（或其他类型的无应答）进行校正的事后分层并不能保证一定能改善估计。但是，如果“追踪调查中的丢失”与在调整程序中用到的变量相关，那么调整后的估计应该比未调整的估计更好。

我们来看某个社区精神健康调查（ECA 调查的其中一个场所）的例子。只有 74.3%的初始受访者在第一轮的年度追踪调查中成功受访。这个损耗程度太大以至于我们没法忽略不计。因此，我们挑选了一些预测变量，对损耗（1＝丢失，0＝受访）进行了 logistic 回归分析，结果如表 3.3 所示。卡方值说明，模型中的变量与损耗显著相关。如表 3.3 所示，效果编码需要每个变量都省略其中一个取值。基于估计到的 beta 系数，对每个受访者都计算了 logit 估计值（$\hat{\lambda}_i$，$i=1, 2, \cdots, n$）。然后这些值被转化成损耗的预测比例 $\hat{p}_i = 1/(1+e^{-\hat{\lambda}_i})$。通过把在追踪调查中丢失的人的权重设为 0，以及把成功受访者

的权重除以$(1-\hat{p}_i)$，这些成功受访者的初始权重增加了。

表 3.3 追踪调查中损耗调整的 logistic 回归模型

Logistic Regression Model						
Factors	*Category*	*Variable*	*Beta Coefficient*	*Survey-Related Information*		
Intercept		-	-0.737*	Initial survey		
Age	18-24 yrs	AGE1	0.196	Design:		Multistage
Sampling						
	25-34 yrs	AGE2	0.052	Sample size:		4967
	35-44 yrs	AGE3	-0.338*	Weighted sum:		300113
	45-54 yrs	AGE4	-0.016			
	55 and over	—		Follow-up survey		
Marital status:	Sep./divorced	MAR	0.051	Sample size:		3690
(74.3%)						
	Other	-		Sum of attrition-		
Gender:	Male	SEX	0.084	adjusted weights:		300172
	Female	—		Adjusted sum:		300113
Race	White	RACE1	-0.668*			
	Black	RACE2	-0.865*	Comparison of attrition-adjusted		
and Other	-			attrition-unadjusted		
estimates						
Socioeconomic	1st quartile	SES1	0.534*			
status	2nd quartile	SES2	0.389*	Disorders	Unadjusted[a]	Adjusted[b]
Diff.						
	3rd and 4th	-				
Family size:	One	SIZE1	0.033	Any disorder	39.2%	43.8%
-4.6%						
	2-4 members	SIZE2	-0.003	Major depre.	15.0	18.1
-3.1						
	5 or more	-		Cog. Impair.	1.8	1.3
0.5						
Diagnosis:	Cog. Impair.	DX1	0.472*	Phobias	8.4	7.6
-0.4						
	Schizophrenia	DX2	-0.049	Alcohol abuse	13.9	13.7
-4.2						
	Antisocial	DX3	0.412*	Schizophrenia	0.6	0.6
-0.4						
	Anorexia	DX4	2.283*			
	No disorder	—	.			

Likelihood ratio chi-square(16) = 198.0, $p < 0.00001$.

注：a. 基于最初的权重。
b. 基于损耗调整后的权重。
* $p < 0.05$.

正如表 3.3 所示，对保留下来的成员调整后的权重总和为 300172，只比初始调查中的权重总和大 59，这个数据意味着这种方法的使用很合理（如果调整后的权重总和与初始调查中的权重总和之间存在很大的差异，调整过程就可能值得担忧）。这些调整后的权重再次被调整以与初始调查达到一致。为了说明使用损耗调整权重的作用，如表 3.3 所示，我们对 6 种精神障碍的流行率分别进行了使用和不使用损耗调整权重这两种不同的估计。我们可以发现，使用了调整权重后，所有障碍（由 DSM-III 定义的障碍）的流行情况比没有调整时的情况大约

高了 5 个百分点。

评估精确度的得失：设计效应

如前一章所述，SRSWOR 样本均值的方差等于 SRSWR 样本均值的方差乘以有限总体校正因素 $(1-f)$。因此，SRSWOR 抽样方差与 SRSWR 抽样方差的比率等于 $(1-f)$，反映了(与使用 SRSWR 相比)使用 SRSWOR 的影响。这个把某个特定抽样设计中某个统计值的方差与 SRSWR 中的进行比较的比率被称为那个统计值的设计效应。它被用于评估与 SRSWR 设计方案相比，现有使用方案中的样本估计其精确度的得失。设计效应小于 1 意味着若要得到与 SRSWR 相同的精确度，需要更少的观察值；设计效应大于 1 意味着要得到与 SRSWR 相同的精确度，需要更多的观察值。在 SRSWOR 设计中，设计效应小于 1，但当抽样比例很小的时候接近于 1。由于研究者习惯上常用 SRSWOR 代替 SRSWR，因此他们倾向于以 SRSWOR 而不是 SRSWR 为基础来计算设计效应。此外，在复杂调查中，设计效应的计算通常基于 SRSWOR 设计中加权后统计值的方差。我们在后文中也将如此计算。

把设计效应这个概念与样本规模结合起来，有效的样本规模可以定义为实际样本规模除以设计效应。如果某个抽样设计的设计效应大于 1，那么实际上为了统计分析样本规模就会减少，因此导致更大的抽样误差。换言之，当设计效应大于 1 时，有效样本规模小于实际样本规模。

我们来检验一下更加复杂的抽样设计中的设计效应。大家已经知道分层随机抽样中抽样误差的性质和估计情况，以及在哪些条件下分层会产生比 SRS 更小的方差。但是，分层通常与其他的设计特征一起使用，如阶层内不同阶段的集群抽样。如前面讨论的，集群会增加抽样误差。在许多实际的抽样设计中，分层的作用会因集群的作用而减弱。可惜我们无法在理论上分别从分层和集群的特征中评估设计效应，反而必须从数字上大致地判断它们的联系效应。

接下来的例子说明了在一个相对简单的情况下设计效应的计算。考虑一个单步骤集群抽样，其中所有的集群大小规模一样。假设在一个高中学校有 N

个英语班级，每个班级有 M 个学生。从这 N 个班级中用 SRS 选择 n 个班级，然后要求这 n 个被选中的班级里的所有学生都报告自今年年初以来他们读过的书的数量。在总体中学生的数量为 NM，在样本中学生的数量为 nM。抽样比例为 $f = nM/NM = n/N$。

因为班级大小相同，平均每个学生所读的书的数量（总体均值）等于 N 个班级均值的平均数。这 n 个样本班级可以看做具 N 个均值的总体的一个随机样本，这个随机样本具有 n 个均值。因此样本均值($\bar{\bar{y}}$)对总体均值($\bar{\bar{Y}}$)来说是无偏的，而它的方差，应用方程 2.1 等于

$$\hat{v}(\bar{\bar{y}}) = \frac{s_b^2}{n}(1-f), \tag{3.3}$$

其中，$s_b^2 = \sum(\bar{y}_i - \bar{\bar{y}})^2/(n-1)$，是集群均值的方差估计值。或者，方程 3.3 可以依据估计到的 ICC($\hat{\rho}$)写成(Cochran，1977，第 9 章)：

$$\hat{v}(\bar{\bar{y}}) = \frac{s^2[1+(M-1)\hat{\rho}]}{nM}(1-f), \tag{3.4}$$

其中 $s^2 = \sum\sum(y_{ij} - \bar{\bar{y}})^2/(nM-1)$，是样本成员的方差。如果方程 3.4 除以 SRSWOR 样本（样本规模为 nM）的均值方差 $\left[\hat{v}(\bar{\bar{y}}) = \frac{s^2}{nM}(1-f)\text{，方程2.1}\right]$，那么这个集群抽样的设计效应为 $1+(M-1)\hat{\rho}$。

当 ICC $= 0$ 时，设计效应为 1；当 ICC > 0 时，设计效应大于 1。如果集群的形成是随机的，那么 ICC $= 0$；当每个集群内部所有成员有相同的取值时，ICC $= 1$。社区调查中用到的大部分集群由相同地区的住宅构成，这样对该调查中许多变量产生的 ICC 是正数。通常来说，社会经济变量的 ICC 比人口特征变量如年龄性别等的 ICC 大。

对更加复杂的抽样设计的设计效应进行评估并不是一个用统计教科书中的公式就可以完成的例行任务；相反，它需要特殊的技术，而这些技术要使用一些不常见的策略。下一章将会介绍用于估计复杂调查中统计值的抽样方差和不同调查的设计效应的几种策略。

调查数据分析中抽样权重的使用

如上面所讨论的，抽样权重可用于计算点估计。所有的点估计都表现为加权后的统计值的形式。从上述讨论中可见，抽样权重的这种用法使得抽样权重的创建和调整显得很有道理，尤其是对描述性分析来说。但是，在分析性研究中抽样权重的使用并不如在描述性分析中那样清楚。如第 2 章最后一部分所述，关于调查数据有 n 种不同的视角。从基于设计的角度看，抽样权重的使用在描述性和分析性研究中都非常重要。推论是以对有限总体进行重复抽样为基础的，而用于推论的概率结构是由表明被包含在样本之中的随机变量界定的。但是，从基于模型的视角看，样本选择方案对于在某个具体的模型下进行推论是无关紧要的。如果总体中的观察值确实符合模型，那么只要选择概率只通过模型中的自变量[①]（取决于该模型中的因变量），抽样设计就不会有任何影响。许多学者已经对在哪些条件下做推论时可以忽略抽样方案的作用做了大量的研究（Nordberg，1989；Sugden & Smith，1984）。而且自 20 世纪 70 年代初期起，很多调查统计学家就已经针对这两种不同的视角进行过争论（Brewer，1999；Brewer & Mellor，1973；Graubard & Korn，1996，2002；Hansen，Madow & Tepping，1983；Korn & Graubard，1995a；Pfeffermann，1996；Royall，1970；Sarndal，1978；T. M. F. Smith，1976，1983）。

好的调查数据的分析需要对这两种视角有一个全面的理解，并且还要考虑到某些实际问题，尤其是在社会调查中。基于模型的方法与其他统计分析领域日益增多的模型推论一致，而且确实提供了一些理论优势。基于模型的估计可以用于相对较小的样本之中，甚至可用于非概率样本中。此外，基于模型的分析可以用标准的统计软件如 SAS 和 SPSS 执行，而不需要依赖如 SUDAAN 及其他本书所提及的调查软件包。但是，基于模型的方法假定该模型正确地反映了自然的真实状态。如果模型设定是错的，那么分析就会有偏差，对数据的解

---

① 而不是其他没有被包含在模型中的因素。——译者注

释就会有误。可惜在社会调查中很难对总体中的所有观察值建立用理论推导出来的模型。此外，模型中相关变量的缺失在调查数据的二手分析中也是一个大问题，因为研究者并不能获得所有的相关变量。因此，基于模型的推论其主要的挑战是建立一个符合分析目的的正确的模型。

我们已经知道加权后的分析会受到具有极大权重的观察值的严重影响（极大权重的产生经常来自无应答和事后分层调整，而不是选择概率）。此外，加权的另一个局限在于方差的增加。通常，当权重的变异较大时方差会增加很多。如果实际上并没有减少偏差的必要，那么使用加权后的估计就会损失一些东西，相比未加权分析，加权后的分析也就会变得低效。Korn 和 Graubard(1999，第 4 章)讨论了处理总体参数的加权和未加权估计的不同问题，并提供了一种检验加权估计的低效性的方法。他们建议，如果低效性并不是严重到不可接受，那么就用加权后分析，以避免未加权分析中的偏差；如果低效性太严重以至于不能接受，那么就用未加权分析，并用调查设计变量——包括权重——以充实模型从而减少偏差。但是，把设计变量结合到模型之中通常会出现问题，因为把设计变量作为附加的解释变量放到模型中可能会与科学的分析目标相矛盾。比如，当分析的目标是检验卫生措施与风险系数两者的相关性时，把设计变量放在模型中可能会干扰这两者的相关路径。

在复杂的大型调查中，通常不可能把所有的设计信息包含在模型之中，尤其在对抽样权重进行了无应答和事后分层调整后(Alexander，1987)。把设计变量放到模型中的方法还有另一个操作性问题，那就是数据常常没有相关的信息。并不是所有与设计相关的信息研究者都能得到。大部分公开的调查数据只提供初级抽样单位(PSU)，而没有次级集群单位（如普查小区或者电话交换台）。出于保密性考虑，他们通常不可能提供次级抽样单位的信息。

在基于模型的分析中，我们必须避免可能的模型设定错误和相关解释变量的缺失。抽样权重的使用（基于设计的分析）可为此提供避免错误的模型设定(DuMouchel & Duncan，1983；Pfeffermann & Homes，1985)。Kott(1991)指出，在线性回归中需要使用抽样权重，因为在大部分二手分析中对调查数据解释变量的选择具有局限性。本书第 6 章最后一节将对使用抽样权重的利弊做进一步讨论。

第 4 章 | 方差估计的策略

对调查统计值进行方差估计是很复杂的，不仅由于抽样设计的复杂性（如前面章节所述），也由于统计值的形式。即使是在 SRS 设计中，某些统计值的方差估计也需要非标准化的估计技术。比如，标准教材明显没有涉及中位数的方差，而且比率估计值的抽样误差（参考注释 1）也很复杂，因为分子和分母同时都是随机变量。某些在标准教科书中没有提及的方差估计方法就有足够的灵活性，可以同时顾及抽样设计的复杂性和统计值的各种不同的形式。本章所介绍的这些常用的方差估计方法包括：复合抽样，对称重复抽样（BRR），“折叠式”重复抽样（JRR），自主抽样法和泰勒级数法。

复合抽样：一种通用的方法

这种方法的根本原理是通过选取一系列复合子样本而不是单个样本来辅助方差计算。它要求每个子样本都用同一个选择方案独立地被选中。然后用同样的方法对每一个子样本进行估计，这样我们就可以通过这些独立的子样本估计值的变异情况，对基于所有子样本的总体估计的抽样方差进行估计。这与第 2 章中重复系统抽样的原理一样。

对于参数 U 的 t 个复合样本估计值 u_1，u_2，…，u_t 的均值（$\bar{u}$），其抽样方差可以通过下面这个简单的方差方程估计出来（Kalton，1983：51）

$$v(\bar{u}) = \sum (u_i - \bar{u})^2 / t(t-1) \qquad [4.1]$$

这个估计可以用于所有抽样设计为独立复合样本得到的所有样本统

计值。

在应用这个估计方法时，Deming(1960)建议对描述性统计值使用10个复合样本，其他人(Sudman，1976)则建议最少用4个。当复合样本的数量在3—13之间时，标准误差大致的估计值可以计算为等于复合样本估计值的幅度除以复合样本的数量(Kish，1965：620)。但是，因为在统计推论中，t个复合样本的方差估计基于$t-1$个自由度，因此在分析性研究中我们需要更多的复合样本，大约在20—30之间(Kalton，1983：52)。

为了充分理解复合抽样方案的策略，我们来看一个简单的例子。假设我们想估计200个新生婴儿中男孩的比例。我们用Cochran(1977：19)书中的随机数字模拟这个调查，假设基数代表男孩。从该书表中的前10列，我们选择了10个复合样本，每个样本的样本规模$n=20$。每个复合样本中男孩的数量如下：

每个复合样本中男孩的数量	9	8	13	12	14	8	10	7	10	8	总计＝99
男孩的比例	0.45	0.40	0.65	0.60	0.70	0.40	0.50	0.35	0.50	0.40	比例＝0.495

男孩的总体百分比为49.5%，其标准误差为$\sqrt{49.5\times 50.5/200}=3.54\%$。用方程4.1从这10个复合样本估计值中估计出来的标准误差为3.58%。要得到一个约为3.50%的大致估计很简单，只要用这些估计值的幅度(70%－35%)除以复合样本的数量10即可。复合抽样方法的主要优势在于很容易对标准误差进行估计。

在实际中，选择独立复合样本的根本原则在某种程度上并不严格。例如，复合样本的选择使用的是无放回抽样而不是有放回抽样。对于不等概率设计，基本权重的计算以及无应答和事后分层的调整通常只对全样本执行一次，而不会对每一个复合子样本都分别进行这些计算和调整。在集群抽样中，研究者通常用与集群最初被选中时相同的次序系统地把不同的集群分配给t个复合样本，这样可以充分利用分层效应。在使用方程4.1时，从全样本

中得到的样本均值通常用做复合样本均值的平均数。这些偏离根本原则的做法会影响方差估计,但在大型调查中,这种偏差应该是可以忽略的(Wolter, 1985: 83—85)。

作为美国国立精神卫生研究院 1984 年 ECA 调查的一部分,在康涅狄格州纽黑文县举行的社区精神卫生调查(E. S. Lee, Forthofer, Holzer & Taube, 1986)是复合抽样的一个典型的例子。这个调查的抽样框是按地理位置排列的居住电网清单。这个调查把两个居住单位作为一个集群,随机选取起始点,以 61 户为间距进行系统抽样。样本中的这一连串集群随后按顺序分配给 12 个子样本。这些子样本用于在长时间的筛选和访问中辅助数据的调度和中期分析。其中 10 个子样本用于社区调查,其他 2 个用于其他分析。这 10 个复合样本在此用于说明方差估计的程序。

这些子样本并没有严格遵循独立复合抽样的基本原则,因为除了第一个随机的起始点,其他起始点都是系统地选取的。但是,在这个例子中,系统地把集群分配给子样本可以产生一个近似的分层,从而得到更加稳定的方差估计,也因此比对这些相对少量的复合样本使用随机起始点更加可取。

所以,我们把这些子样本当做复合样本,并用方程 4.1 的复合抽样方差估计法进行估计。

通过 Kish 随机选择表(Kish, 1949),每一个家庭里面有一个成年人被随机选中,因此每个家庭里面成年人的数量就是每个观察值的样本权重。对这个权重随后进行无应答和事后分层调整。样本权重的创建只对全样本而不会对每一个子样本分别进行。这即为在分析中使用的权重。

表 4.1 展示了对全样本以及各复合子样本计算出来的三种统计值。(精神障碍)流行率的方差估计值,是百分数 (p) 的形式,可以利用方程 4.1 从复合样本估计值(p_i) 中计算出来:

$$v(p) = \frac{\sum (p_i - 17.17)^2}{10(10-1)} = 0.3474,$$

表 4.1 复合样本中标准误差的估计:纽黑文县 ECA 调查,1984($n = 3058$)

			回归置信区间[a]			
复合样本	流行率[b]	数 比[c]	截 距	性 别	种 族	年 龄
全样本	17.17	0.990	0.2237	−0.0081	0.0185	−0.0020
1	12.81	0.826	0.2114	0.0228	0.0155	−0.0020
2	17.37	0.844	0.2581	0.0220	0.0113	−0.0027
3	17.87	1.057	0.2426	−0.0005	0.0393	0.0015
4	17.64	0.638	0.1894	0.0600	0.2842	−0.0029
5	16.65	0.728	0.1499	0.0448	−0.0242	−0.0012
6	18.17	1.027	0.2078	−0.0024	−0.0030	−0.0005
7	14.69	1.598	0.3528	−0.0487	−0.0860	−0.0028
8	17.93	1.300	0.3736	−0.0333	−0.0629	−0.0032
9	17.86	0.923	0.2328	−0.0038	0.0751	−0.0015
10	18.91	1.111	0.3008	−0.0007	0.0660	−0.0043
范围	6.10	0.960	0.2237	0.1087	0.3702	0.0038
标准误差						
复合样本	0.59	0.090	0.0234	0.0104	0.0324	0.0004
SRS	0.68	0.097	0.0228	0.0141	0.0263	0.0004

资料来源:取自"Complex Survey Date Analysis: Esimation of Standard Errors Using Pseudo-Strata," E. S. Lee, Forthofer, Holzer, and Taube, *Journal of Economic and Social Measurement*, 1986 年版权所有。

注:a. 因变量(1 为状态存在,0 为状态不存在)对性别(男 = 1,女 = 0),肤色(黑人 = 1,非黑人 = 0)和年龄做回归。这个分析只是作为演示而已。

b. 过去 6 个月有任何一种精神疾病的百分比。

c. 6 个月流行率的性别差异。

其标准误差为$\sqrt{0.3474} = 0.59$。由于应答率不同,总体流行率 17.17%与 10 个复合样本估计值的均值稍微不同。我们注意到,复合样本估计值幅度的十分之一①,即 0.61 与从方程 4.1 中得到的标准误差很接近。类似地,还可以对比数比(odds ratio)和回归系数估计标准误差。这些标准误差的估计值与从假定简单随机抽样(用教科书中的合适的方程)而计算出来的值几乎相同。这意味着对这个调查中的这些统计值来说设计效应非常小。

① 幅度除以样本数 10。——译者注

虽然复合抽样设计为我们提供了一种易于计算的方差估计方法，但为了获得合格的统计推论的精确度，还必须要有足够数量的复合样本才可以。然而，如果复合样本的数量很多而每个样本的样本量却相对较小，就会严重限制在每个复合样本里面使用分层。更重要的是，在复杂抽样设计中使用复合抽样是很不切实际的。因此，复合抽样设计很少在大型的分析调查中使用。相反，复合抽样的原理常在数据分析阶段被用于估计方差。这种尝试促使了方差估计的拟复合抽样方法的出现。下面要讲的就是基于这种拟复合抽样的原理而产生的两种方法。

对称重复抽样

对称重复抽样(BRR)是复合抽样原理在配对选择设计中的应用，其中每一个阶层内都有两个初级抽样单位(PSUs)被选取。配对选择设计代表了分层的最大化使用并允许对方差进行计算。在这种情况下，两个抽样单位之间的方差等于他们之间差异的平方的一半。为了应用复合抽样的原理，我们首先把样本分成随机的组别以形成拟复合样本。如果是分层的设计，需要所有的阶层都能在每个拟复合样本中出现。在一个分层的配对选择设计中，我们只能建立两个拟复合样本：第一个从每个阶层的两个抽样单位中选出其中一个；每个阶层中剩下的另一个抽样单位则组成第二个拟复合样本(补充样本)。每个拟复合样本几乎都包含了总样本的一半。

把 $t=2$ 代入方程4.1，我们可以对这两个复合样本估计值 u' 和 u'' 的均值的抽样方差进行估计，通过：

$$v(\bar{u}) = [(u' - \bar{u})^2 + (u'' - \bar{u})^2]/2 \qquad [4.2]$$

如上所示，复合样本估计值的均值经常用从全样本中获得的总体估计值代替。但是这个估计值太不稳定，以至于没有任何的实际价值，因为它只基于两个拟复合样本。BRR 则通过重复建立半样本的复合样本和从不同的阶层中选取不同的抽样单位解决了这个问题。这些拟复合的半样本因此包含了某些共同的单位，这样就导致了复合样本间的相关，从而使估计变得复杂了。其中一

个解决方法——这种方法对线性统计值得出的方差估计是无偏的——是通过使用正交矩阵平衡拟复合样本的构造(Plackett & Burman, 1946)。充分的平衡需要矩阵的大小等于 4 的倍数而且复合样本的数量要大于或等于阶层的数量。这样某个样本统计值的抽样方差就可以估计为方程 4.2 中得到的方差估计值的均值除以 t 个复合样本数:

$$v(\bar{u}) = \sum[(u'_i - \bar{u})^2 + (u''_i - \bar{u})^2]/2t = \sum(u'_i - u''_i)^2/4t \qquad [4.3]$$

通过去掉补充的半样本的复合样本,可以把方程简化为

$$v'(\bar{u}) = \sum(u'_i - \bar{u})^2/t \qquad [4.4]$$

这个估计方法最初是由 McCarthy(1966)提出的。McCarthy 阐释了这种平衡的方法以产生线性估计的无偏方差估计值。对非线性估计来说,方差估计是有偏差的,但很多研究证明这种偏差很小。在阶层数目较大的情况下,通过使用更少量的部分对称复合样本(K. H. Lee, 1972; Wolter, 1985: 125—130),还可以进一步简化这种计算方法。

就像在复合抽样中一样,BRR 假定初级抽样单位(PSU)是在阶层内通过有放回抽样而获得的,虽然实际操作中通常用的是无放回抽样。理论上,当应用到无放回样本中时,会导致方差的过高估计,但这种过高估计其实可以忽略不计,因为在抽样比例很小时,无放回抽样方案选取同一个单位超过 1 次的机会很低。配对选择设计中抽样比例(假定在 BRR 方法中)通常很小,因为从每个阶层内只选取了两个 PSU。

当用于多阶段选择设计中时,BRR 通常只应用到 PSU 而无视 PSU 内部的次级抽样。这是必定的,因为当第一阶段的抽样比例很小时,抽样方差可以充分地从 PSU 之间的变异情况大概估算出来。这被称为最终集群逼近法。正如 Kalton(1983,第 5 章)所描述的,对简单两阶层选择设计的方差的无偏估计包括从这两个阶层中各得到的一个组成部分,但从第二个阶层中得到那个组成部分要乘以第一个阶层的抽样比例。因此,随着第一个阶层抽样比例的减少,第二个阶层的作用可以忽略不计。这种只基于 PSU 的快捷的计算方法在为复

杂数据做公开准备以及对这种数据进行分析时显得尤为方便,因为除了第一阶段的抽样信息,并不需要这些复杂设计的其他详细的特征信息。

如果 BRR 要用于配对选择设计之外的其他方案,就有必要对数据结构进行调整以适应这种方法。在许多多阶段调查中,分层的数目被最大化了,最终只从每个阶层中选取 1 个 PSU。在这种情况下,PSU 之间可以进行配对以形成合并层从而使用 BRR。这种方法通常会导致对方差的某种过高估计,因为某些阶层间的变异性现在被包含在阶层内的计算结果之中了。对线性统计值来说,如果这种合并是明智而审慎地进行的,那么这个问题就不是很严重;但是,对非线性统计值,通常不建议使用这种合并方式(Wolter,1985: 48)。稍后介绍的泰勒级数法可以用于非线性统计值的估计。还有一种对每个阶层内包含的 3 个 PSU 建立正交平衡的方法,但它并没有被广泛使用(Gurney & Jewett, 1975)。

现在我们把 BRR 这种方法应用到 1984 年的美国 GSS。正如前一节所述,它用的是多阶段选择设计。第一阶段的抽样从 84 个阶层——郡级地区或郡级地区群——中各选取一个 PSU。前面的 16 个阶层为大都市地区,被定为具有自我代表性(或者自动包含在样本之中)。为了使用 BRR,这 84 个阶层被合并成 42 对拟阶层。因为数据中不具自我代表性的 PSU 的编号服从阶层的地理排序,因此配对是以 PSU 的编码为基础按顺序进行的。之后,再把这 16 个具自我代表性的阶层合并成 8 个拟阶层,而剩下的 68 个不具自我代表性的阶层则被合并成 34 个阶层。但是,对这些具自我代表性的阶层进行的配对方式错误地包括了它们之间的异质性。为了排除这些,而仅仅只包括每个具自我代表性的阶层内部的异质性,这些在每个具自我代表性的拟阶层内部的观察值被结合起来再随机地分成了两个拟初级抽样单位。

为了平衡从 42 个拟阶层中产生的半样本的复合样本,我们使用了一个 44 阶正交矩阵(见表 4.2)。这个矩阵由数字 0 和 1 构成,为了和这 42 个阶层匹配,前面 2 列被舍弃了(即 44 行对应复合样本,42 列对应拟阶层)。0 表示包括阶层中的第一个 PSU,1 表示包含第二个 PSU。行是复合样本,列是阶层。比如,第一个复合样本包含了 42 个拟阶层中每个阶层的第二个 PSU(因为第 1 行所有的值都等于 1)。利用正交矩阵的行,44 个复合样本和 44 个补充复合样本就产生了。

表 4.2　44 阶正交矩阵

行	列(44)
1	11111111111111111111111111111111111111111111
2	10100101001110111110001011100000100011010110
3	10010010100111011111000101110000010001101011
4	11001001010011101111100010111000001000110101
5	11100100101001110111110001011100000100011010
6	10110010010100111011111000101110000010001101
7	11011001001010011101111100010111000001000110
8	10101100100101001110111110001011100000100011
9	11010110010010100111011111000101110000010001
10	11101011001001010011101111100010111000001000
11	10110101100100101001110111110001011100000100
12	10011010110010010100111011111000101110000010
13	10001101011001001010011101111100010111000001
14	11000110101100100101001110111110001011100000
15	10100011010110010010100111011111000101110000
16	10010001101011001001010011101111100010111000
17	10001000110101100100101001110111110001011100
18	10000100011010110010010100111011111000101110
19	10000010001101011001001010011101111100010111
20	11000001000110101100100101001110111110000011
21	11100000100011010110010010100111011111000101
22	11110000010001101011001001010011101111100010
23	10111000001000110101100100101001110111110001
24	11011100000100011010110010010100111011111000
25	10101110000010001101011001001010011101111100
26	10010111000001000110101100100101001110111110
27	10001011100000100011000110010010100111011111
28	11000101110000010001101011001001010011101111
29	11100010111000001000110101100100101001110111
30	11110001011100000100011010110010010100111011
31	11111000101110000010001101011001001010011101
32	11111100010111000001000110101100100101001110
33	10111110001011100000000011010110010010100111
34	11011111000101110000010001101011001001010011
35	11101111100010111000001000110101100100101001
36	11110111110001011100000100011010110010010100
37	10111011111000101110000010001101011001001010
38	10011101111100010111000001000110101100100101
39	11001110111110001011100000100011010110010010
40	10100111011111000101110000010001101011001001
41	11010011101111100010111000001000110101100100
42	10101001110111110001011100000100011010110010
43	10010100111011111000101110000010001101011001
44	11001010011101111100010111000001000110101100

资料来源:经出版社同意,取自 Wolter(1985,第 32 题)。

为了对从全样本中产生的统计值进行方差估计，我们首先需要对这 44 个复合样本和补充复合样本分别计算出我们关注的统计值。在计算这些复合样本的估计值时，我们使用了调整后的抽样权重。表 4.3 显示了对这 44 个复合样本

表 4.3　BRR 复合样本中同意一个成年人攻击另一个成年人的比例估计值：美国综合社会调查，1984（ $n=1473$ ）

重复次数	估计值(%)		重复次数	估计值(%)	
	复合样本	补充复合样本		复合样本	补充复合样本
1	60.9	59.2	23	61.4	58.6
2	60.1	59.9	24	57.7	62.4
3	62.4	57.9	25	60.4	59.6
4	58.5	61.7	26	61.7	58.2
5	59.0	61.0	27	59.3	60.6
6	59.8	60.2	28	62.4	57.6
7	58.8	61.5	29	61.0	58.9
8	59.0	61.0	30	61.2	58.7
9	61.3	58.8	31	60.9	59.1
10	59.2	60.8	32	61.6	58.5
11	61.7	58.3	33	61.8	58.2
12	60.2	59.8	34	60.6	59.4
13	62.1	58.7	35	58.6	61.5
14	59.7	60.4	36	59.4	60.7
15	58.1	62.0	37	59.8	60.3
16	56.0	64.2	38	62.0	58.1
17	59.8	60.3	39	58.1	61.9
18	58.6	61.3	40	59.6	60.5
19	58.6	61.1	41	58.8	61.2
20	60.8	59.3	42	59.3	60.8
21	63.4	56.5	43	58.7	61.4
22	58.3	61.7	44	60.5	59.5

总体估计 = 60.0

方程估计	方差	标准误	设计效应
基于方程 4.3	0.000231	0.0152	1.42
基于方程 4.4	0.000227	0.0151	1.40

和他们的补充复合样本分别估计出的同意“攻击”的成年人的比例。总体比例为 60.0%。由方程 4.3 估计出来的总体比例的抽样方差为 0.000231。把这个结果与 SRS 设计中该比例的抽样方差 [$pq/(n-1)=0.000163$，忽略 FPC]进行比较，可知设计效应等于 1.42($=0.000231/0.000163$)。这个设计效应意味着从 GSS 中得到的比例估计值的方差比从具相同样本规模的 SRS 中得到的方差大 42%。从方程 4.4 得到的方差估计值也类似。

综上所述，BRR 使用了拟复合抽样的方法来估计抽样方差，它主要是为配对选择方案而设计的。它也可以用于某些通过配对阶层而只从每个阶层中选择一个 PSU 的复杂的调查，但是这种配对方法必须是明智而审慎的，而且要考虑到实际的样本选择程序。在大多数可用的 BRR 的应用软件包中，对于复合样本中被选中的 PSU，其观察值的抽样权重被增加至两倍，以弥补没有被选中的那一半 PSU。此外，Fay 认为，在创建复合权重时还存在一个 BRR 的变量，用 $2-k$ 还是 k 乘以初始权重，主要取决于以正交矩阵($0\leqslant k<1$)为基础的 PSU 是被选中了还是没有被选中。这将在下一章进一步说明。

“折叠式”重复抽样

“折叠式”方法最初作为一种估计偏差的非参数程序，是由 Quenouille (1949)提出的，随后 Tukey(1958)介绍了相同的这种程序怎样用于估计方差。Durbin(1959)首先在他的开创性比率估计研究中使用了这种方法。之后，Frankel(1971)采用与 BRR 相同的方式，把它应用于复杂调查中的方差计算当中，而且还把这种方法命名为“折叠式”重复抽样法。跟 BRR 一样，JRR 技术通常用于阶层内的 PSU。

我们可以通过对一个简单随机样本的样本均值进行抽样方差估计来阐释“折叠式”方法的基本原理。假设 $n=5$，y 的样本取值为 3，5，2，1 和 4。那么样本均值 $\bar{y}=3$，它的抽样方差(忽略 FPC)等于

$$v(\bar{y})=\frac{\sum(y_i-\bar{y})^2}{n(n-1)}=0.5 \qquad [4.5]$$

均值的“折叠式”方差的产生如下：

(1) 删除第一个样本值，然后计算出拟样本均值，即 $\bar{y}_{(1)}=(5+2+1+4)/4=12/4=3$。接着，换成删除第二个样本值，得到第二个拟均值 $\bar{y}_{(2)}=10/4$；以此类推，$\bar{y}_{(3)}=13/4$，$\bar{y}_{(4)}=14/4$，$\bar{y}_{(5)}=11/4$。

(2) 对这 5 个拟均值计算它们的均值 $\bar{\bar{y}}=\sum\bar{y}_{(i)}/n=(60/4)/5=3$，这个均值与样本均值相等。

(3) 然后方差就可以从这 5 个拟均值(每个都包含 4 个观察值)的变异性中估计出来：

$$v(\bar{\bar{y}})=\frac{(n-1)\sum(\bar{y}_{(i)}-\bar{\bar{y}})^2}{n}=0.5 \qquad [4.6]$$

这个结果与方程 4.5 的结果相同。

这些基于复合抽样的程序有一个明显的优势：他们可以同时应用于那些无法用方程表达出来的估计值，如样本中位数，以及那些以公式为基础的估计值。中位数的抽样方差没有可用的公式，但是“折叠式”程序却可以为其提供一个估计值。用上面这个例子，样本中位数是 3，而 5 个对应的拟中位数分别为 3，2.5，3.5，3.5 和 2.5(这些拟中位数的均值为 3)。用方程 4.6，中位数的方差估计为 0.8。

以同样的方式，“折叠式”程序还可以应用于复合抽样。我们可以每次去掉其中一个复合样本并计算出相应的拟值来估计“折叠式”方差，虽然在这种情况下它并没有任何计算优势。但是它也可以应用于任意由概率抽样构成的随机组别。比如，为使用“折叠式”程序，一个系统抽样样本可以被分成随机或系统的子样本。对其他抽样设计来说，可以按照 Wolter(1985：31—33)提出的实际规则构成随机组别。基本原理是，随机组别以下述方式产生：每个随机小组都具有与母样本相同的抽样设计。这需要实际抽样设计的详细的信息，但这样的信息在公开的调查数据中通常是得不到的。因此，“折叠式”程序通常用于 PSU 而不是随机组别。

对于配对选择设计，复合样本是通过——从一个阶层中去掉 1 个 PSU，然

后再对剩下的那个 PSU 加权以保持这个阶层在总样本中的比例——这种方式构成的。补充复合样本的构成方式是一样的，只需要把这个阶层中去掉的和保留的 PSU 对换一下，对每个复合样本都估计一个拟值。对于加权后的样本，保留下来的 PSU 的抽样权重需要加大以弥补那些被去掉的 PSU 里面的观察值。放大后的权重等于保留下来的 PSU 中所有权重的和除以一个因子 $(1-w_d/w_t)$，其中 w_d 是被去掉的 PSU 中所有权重的和，而 w_t 是该阶层内所有 PSU 的权重的和。这个因子代表了被删除的 PSU 占总阶层权重的比例的补集①。然后，在配对选择设计中某个样本统计值的方差就可以由阶层 h 中拟值 u'_h 和补充拟值 u''_h 通过下列方程计算出来：

$$v(\bar{u}) = \sum[(u'_h - \bar{u})^2 + (u''_h - \bar{u})^2]/2 = \sum(u'_h - u'')^2/4 \qquad [4.7]$$

这个估计值与方程 4.3 的形式一样，可以调整为每个阶层只包含一个复合样本(不与补充样本结合起来取均值)，像 BRR 中方程 4.4 一样，得出

$$v'(\bar{u}) = \sum(u'_h - \bar{u})^2 \qquad [4.8]$$

JRR 并不局限于配对选择设计，它可用于每个阶层含任意数量 PSU 的情况。如果我们用 u_{hi} 表示第 h 个阶层第 i 个复合样本的 U 的估计值，n_h 表示第 h 个阶层中被抽中的 PSU 的数量，r_h 表示在第 h 个阶层中形成复合样本的数量，那么方差就可以通过下面的方程估计而得

$$v(\bar{u}) = \sum_h^L \left(\frac{n_h - 1}{r_h}\right) \sum_i^{r_h} (u_{hi} - \bar{u})^2 \qquad [4.9]$$

如果阶层 h 内每个 PSU 都依次被删掉以形成不同的复合样本，在每个阶层内 $r_h = n_h$，但并不需要在第 h 个阶层构造 n_h 个复合样本。当阶层的数量很大而 n_h 大于或等于 2 时，计算方程可以简化为只使用每个阶层内的 1 个复合样本。但是，在分析性研究中，必须使用足够多的复合样本以保证有足够的自由度。

---

① 即 1 减被删除的 PSU 占总阶层权重的比例。——译者注

表 4.4 给出了把 JRR 应用到在 BRR 计算中使用过的 1984 年美国 GSS 合成配对设计中的结果。对 42 个“折叠式”复合样本和他们的补充样本，分别列出了同意“攻击其他成年人”的成年人比例的估计值。利用方程 4.7，我们得出方差估计值为 0.000238，设计效应为 1.46，这个结果与用 BRR 得出的结果差不多。把补充样本去掉只包含 42 个复合样本时（方程 4.8），方差估计值等于 0.000275，设计效应为 1.68。

表 4.4　JRR 复合样本中同意一个成人攻击另一个成人的比例估计值：美国综合社会调查，1984

	估计值(%)			估计值(%)	
重复次数	复合样本	补充复合样本	重复次数	复合样本	补充复合样本
1	60.2	59.8	22	60.3	60.0
2	60.2	59.8	23	60.0	60.0
3	60.0	60.0	24	60.4	59.6
4	60.3	59.8	25	60.1	59.8
5	60.0	60.1	26	59.8	60.3
6	59.9	60.1	27	59.9	60.1
7	60.0	60.0	28	60.1	60.0
8	60.0	60.0	29	59.5	60.3
9	59.9	60.2	30	59.9	60.1
10	60.1	60.0	31	59.6	60.2
11	59.8	60.2	32	60.5	59.6
12	59.9	60.1	33	60.1	59.9
13	59.8	60.2	34	60.3	59.8
14	60.0	60.1	35	60.1	59.8
15	59.6	60.5	36	60.2	59.8
16	60.4	59.6	37	60.0	60.0
17	59.9	60.0	38	59.6	60.4
18	59.8	60.2	39	59.9	60.1
19	59.8	60.2	40	60.5	59.6
20	59.9	60.1	41	60.4	59.8
21	60.0	60.0	42	60.7	59.4

总体估计 = 60.0

方差估计	方差	标准误	设计效应
基于方程 4.7	0.000238	0.0152	1.46
基于方程 4.8	0.000275	0.0166	1.68

仔细观察表 4.4,我们可能会觉得 JRR 复合估计值的变异程度没有表 4.3 中 BRR 复合估计值的变异程度那么大。但我们要注意,JRR 代表一种使用了不同方法估计方差的不同的策略。注意 BRR 中方程 4.3 的分母包含了复合样本的数量 t,而 JRR 中方程 4.7 则与复合样本的数量无关。原因在于,在 JRR 中,复合估计值本身已经与复合样本的数量相关。由于那些复合样本的产生过程已经删除了其中一个单位,因此与只有少量的抽样单位可用于构造复合样本的情况相比,当可供利用的抽样单位数目很大时,得出的复合估计值会更加接近于总体估计值。因此,不需要在方程 4.7 和 4.8 中加入复合样本的数量。但是,当复合样本的数量比 PSU 的总数量少的时候,就需要把复合样本的数量考虑进来,如方程 4.9 所示。

综上所述,JRR 技术以拟复合抽样的方法为基础,可以用于估计复杂抽样调查中的抽样方差。不需要对样本选择方案做任何限制,但产生复合样本时要非常小心,必须把最初的抽样设计考虑在内。正如前面提到的,这些详细的设计信息对二手分析研究者来说很难得到。比如,如果 GSS 数据文件可以提供更多的关于终端集群的信息,我们就可以构造更加方便而且更接近真实抽样设计的随机组群,而不需要把 JRR 技术用于折叠配对设计中。

自主抽样法

与 BRR 和 JRR 密切相关的另一种方法是由 Efron(1979)推广的自主抽样法。其基本原理是通过重复地从观察到的数据中抽取初级抽样单位(PSU)来建立具相同规模且与设计中的结构相同的复合样本。把自主抽样法应用到 GSS 数据中的 42 个拟阶层中的 84 个 PSU 中,我们会抽到 84 个 PSU(使用有放回抽样方法),每两个从同一个阶层中得来。在某些阶层中,相同的 PSU 可能会被选择两次。这种抽样方法常常要重复多次,最少 200(标记为 B)次(Efron & Tibshirani, 1993,第 6.4 节)。但是,如要对更少的变量进行估计,通常就需要多非常多的复合样本(Korn & Graubard, 1999: 33),对每个新建立的复合样本(u_i)都进行参数估计。然后,所有复合样本估计值的均值,其方差的自主抽样估

计值表示为：

$$v(\bar{u})=\frac{1}{B}\sum_{i=1}^{B}(u_i'-\bar{u})^2 \qquad [4.10]$$

我们需要把这个估计值乘以 $(n-1)/n$ 以纠正偏差。当 n 很小时，偏差可能非常大。在我们的例子中，每个阶层内有 2 个 PSU，估计到的方差需要平分。另一个纠正偏差的方法是在阶层 h 内重新抽取（n_h-1）个 PSU，然后把这些抽中的 PSU 内的观测值的抽样权重乘以 $n_h/(n_h-1)$（Efron，1982：62—63)。在我们的例子中，这样会产生如 BRR 中的半样本复合样本。基于至少 200 个复合样本的自主抽样估计与基于 44 个半样本复合样本的 BBR 估计差不多相同。由于自主抽样法需要大量的复制，因此这种方法并没有广泛应用于复杂调查分析的方差估计。

学者们已经提出了把自主抽样法应用到方差估计中的各种不同的方法（Kovar，Rao & Wu，1988)。虽然基本的原则广为人知，但在选择自主抽样样本时却有许多不同的方法。比如，Chao 和 Lo(1985)建议把主样本中的每个观察值复制 N/n 次以产生无放回简单随机抽样的自主抽样总体。对那些不等概率抽样方案来说，观察值的复制需要与抽样权重成比例；也就是说，要用 PPS 选择自主抽样样本。目前，还没有详尽的研究分析这些可供选择的方法以及可能出现的因为偏离独立同分布样本这个基本假设而带来的影响。

虽然自主抽样法对处理许多统计问题非常有用，但对于复杂调查中的方差估计来说，它不如 BRR 和 JRR 实用，因为它需要大量的复合样本。对不同的使用者，BRR 和 JRR 会产生相同的结果，但自主抽样法可能会因不同的使用者或者相同使用者的不同使用方式而出现不同的结果，因为复制程序就很可能每次都产生不同的结果。正如 Tukey(1986：73)所述，与自主抽样法和其他模拟方法相比，“目前折叠法似乎是最实际可行的用于估计众多不确定性来源的方法”。目前自主抽样法并没有被安装在现有的复杂调查分析的软件包中，尽管它已被广泛用于统计计算的其他领域。

泰勒级数法（线性化）

泰勒级数扩展法已经应用于数学和统计的许多不同情况。它的一个早期的应用是用于获取难以计算的函数值的近似值，比如指数 e^x 或者对数[log(x)]函数。它出现在计算机能完成该项特殊功能之前，以及在我们还无法获得相关的适用表时。e^x 的泰勒级数扩展涉及到对 e^x 求关于 x 的一阶导数或者更高阶的导数；对某些值——通常是 0——估计这些导数；并在这些导数的基础上建立一些列（数列/方程）项。e^x 的扩展为

$$1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+\cdots$$

下面这个是以 a 扩展的一般方程的应用：

$$f(x)=f(a)+f'(a)(x-a)+\frac{f''(a)(x-a)^2}{2!}+\frac{f'''(a)(x-a)^3}{3!}+\cdots$$

在统计上，泰勒级数法用于获取某些非线性函数的近似值。通常，近似值为函数提供了合理的估计，有时候近似值甚至是一个线性函数。这种方差估计的方法在现有文献中有几种不同的名称，包括线性化方法、delta 方法（Kalton，1983：44）和方差传递法（Kish，1965：583）。

在统计应用中，扩展估计用的是 x 的均值或者期望值 $E(x)$。如果我们用 $E(x)$ 代替上面那个一般公式中的 a，会有

$$f(x)=f[E(x)]+f'[E(x)][x-E(x)]+f''[E(x)][x-E(x)]^2/2!+\cdots$$

根据定义，$f(x)$ 的方差 $V[f(x)]=E[f^2(x)]-E^2[f(x)]$，使用泰勒级数扩展法，会有

$$V[f(x)]=\{f'[E(x)]\}^2V(x)+\cdots \qquad [4.11]$$

对不止一个随机变量的函数，方法也一样。对具两个变量的函数，泰勒级数扩展法会产生

$$V[f(x_1, x_2)] \cong \left(\frac{\partial f}{\partial x_1}\right)\left(\frac{\partial f}{\partial x_2}\right)Cov(x_1, x_2) \qquad [4.12]$$

把方程 4.12 应用到两个变量 x 和 y 的比——即 $r = y/x$ ——之中，我们会得到比率估计的方差方程：

$$V(r) = \frac{V(y) + r^2V(x) - 2rCov(x, y)}{x^2} + \cdots$$

$$= r^2\left(\frac{V(y)}{y^2} + \frac{V(x)}{x^2} - \frac{2Cov(x, y)}{xy}\right) + \cdots$$

把方程 4.12 扩展到 c 个随机变量的情况，$\theta = f(x_1, x_2, \cdots, x_c)$ 的近似方差为

$$V(\theta) \cong \sum\sum\left(\frac{\partial f}{\partial x_i}\right)\left(\frac{\partial f}{\partial x_j}\right)Cov(x_i, x_j) \qquad [4.13]$$

表 4.5　用泰勒级数法估计的标准误差：同意成年人攻击另一个人的百分比，美国 1984 年综合社会调查（$n = 1473$）

	子样本	估计值(%)	标准误差(%)	设计效应
总体		60.0	1.52	1.41
性别	男	63.5	2.29	1.58
	女	56.8	1.96	1.21
种族	白人	63.3	1.61	1.43
	非白人	39.1	3.93	1.30
教育	大学	68.7	2.80	1.06
	高中毕业	63.3	2.14	1.55
	其他	46.8	2.85	1.27

把方程 4.13 应用到 n 个观察值样本中的 c 个变量的加权后估计中，

$$f(Y) = \hat{Y}_i = \sum w_i y_{ij}, \ j = 1, 2, \cdots, c,$$

Woodruff(1971)指出

$$V(\theta) \cong V\left[\sum w_i \sum\left(\frac{\partial f}{\partial y_j}\right)y_{ij}\right] \qquad [4.14]$$

这种非线性估计的另一种线性方差形式具有计算上的优越性，因为它避免了方程 4.13 中 $c \times c$ 协方差矩阵的计算。一个简单的加法的互换就实现了把多阶段估计问题转化成单变量问题的便利。这种通用的计算程序可用于不同的非线性估计，包括回归系数(Fuller，1975；Tepping，1968)。

在复杂调查中，这种近似值法被应用到阶层内的 PSU 总体。即，方差估计值是同一阶层内不同 PSU 的(方程 4.14 中)方差的加权联合值。这些方程很复杂，但需要的计算时间远比上面我们讨论过的复制方法所用的时间少得多。这种方法可用于任何一种用数学方式表达的统计量，比如均值或者回归系数，但不能用于那些非函数式的统计量，如中位数或者其他百分位数。

我们现在来看 GSS 例子中样本比例的方差估计。表 4.5 列出了按性别、种族和教育程度分析，泰勒级数法应用于同意攻击别人的成年人的百分比时的结果。该比例等于所有正面回复的权重之和与所有权重的比。其标准误差的计算根据方程 4.14，做了相应修改以包括 PSU 和阶层。总体比例的设计效应为 1.41，这与用另外两种方法估计出来的结果(见表 4.3[BRR]和表 4.4[JRR])大体相同。估计到的比例随性别、种族和教育程度的不同而不同。由于子样本很小，所以子样本中的标准误差比总体中的要大。此外，子样本中的设计效应也与总体中的不一样。

这一章中，我们介绍了几种复杂调查中统计量的方差估计方法(更深入的介绍请参考 Rust & Rao，1996)。GSS 和其他调查的例子表明，在大部分复杂调查中，设计效应大于 1。更多的例子，可以参考 E. S. Lee，Forthofer 和Lorimor(1986)以及 Eltinge，Parsons 和 Jang(1997)。第 6 章里的例子也将说明在复杂调查数据分析中上述其中一种方法的重要性。

第 5 章 | 调查数据分析的准备

前几章重点关注的是调查设计的复杂性和这些设计的方差估计方法。在应用抽样权重和估计设计效应的方法之前，我们必须了解调查设计和对统计量进行相关估计的数据要求以及需要使用的软件。相比从 SRS 中得到的数据，对复杂调查数据来说这些要求在某种程度上更加严格。

调查分析的数据要求

正如第 3 章所述，权重和设计效应是恰当的调查数据分析所需的基本因素。在对从二手来源中得到的调查数据做分析准备工作时，除了感兴趣的变量，我们还必须把权重、(数据文件中的)抽样单位和阶层的识别变量也包括进来。由于在不同的调查数据源中这些与设计相关的条目其命名各不相同，因此很有必要仔细阅读相关文件或者向数据来源机构或个人咨询，以了解调查设计和数据准备程序。

在主要的调查数据来源中，权重通常是公开可用的。正如前面指出的，权重反映了选择概率以及无应答和事后分层调整。这些权重通常表示为扩展权重，扩展权重总和等于总体规模。在某些特定的分析中，把它们转化成相对权重后可能更加方便，相对权重总和等于样本规模。某些调查数据可能包含几个权重变量以辅助数据的恰当使用，这几个权重可能对应于不同的目标总体或者特定数据条目的子样本。通过仔细阅读相关文件，对不同的数据用途选择适当的权重是很有必要的。对某些调查来说，权重这个变量并没有明确标明，因此需要研究抽样设计以产生这些权重。如第 4 章所述，在 GSS 中，权重是从家庭

中成年人的数量得出的。此外，也很有必要进行事后分层调整以使样本构成与总体构成一致(如表 3.1)。如果在与数据提供方联系后仍然无法得到权重，使用者也不能假定数据具有自加权性质。如果没有权重就用数据，使用者必须在报告结果时明确说明这点。很难想象在没有权重的情况下对调查数据进行分析，即便是在基于模型的分析之中，这种情况我们也会意识到可能存在不等选择概率和子样本的不同应答率(虽然与基于设计的分析使用方法不同)。

设计效应的计算通常需要第一阶段选择程序方面的信息，即，层和 PSU 的识别；在某些特定的嵌入式设计中，还需要次级抽样单位和相关的层方面的信息。如果从每个层内选择一个 PSU(如在 GSS 中)，层的标识与 PSU 的标识是一样的。如果没有用到分层或者数据并没有提供层标识，那么我们可以假定其为无限制集群抽样，然后进行分析。如果没有任何阶层与 PSU 信息，那么就很有必要分析一下，对于给定的抽样设计，把数据当成 SRS 是否合理。

当使用了分层且层标识可供利用时，我们需要确保每个阶层内至少有 2 个 PSU，否则就不可能估计方差。如果每个阶层内只有一个集群被选中，就需要对层进行配对以产生拟阶层。对层进行配对需要充分理解抽样设计。第 3 章列举了 GSS 中一个特定的阶层配对策略的例子。当相关文档中没有任何有用的信息时，也可使用随机匹配的方法。基于美国国民健康调查，Stanek 和 Lemeshow(1977)研究了匹配的效应，发现加权后均值的方差估计和联合比估计对不同的阶层匹配方法没有什么反应，但这个结果可能并不是对所有的调查都适用。

预备性分析的重要性

调查数据分析以前期探索研究为起点，以检验数据是否适合用于有意义的分析。对二手数据来源的前期研究中，一个重要的考虑因素是要检验不同的子群体中是否存在足够的观察值以支持后面的分析。基于未加权的列表结果，分析者可以确定样本规模是否足够大，以及变量的不同类别是否需要合并。未加权的列表结果同时还能显示具缺失值或者极端值的观察值的数量，可反映是否

存在测量误差或者录入错误。

虽然单元无应答调整是由数据收集机构在创建抽样权重时进行处理的，但分析者必须自己处理缺失值，即选项无应答。如果缺失值较少，我们可以在分析中忽略那些具缺失值的受访者。但是，基于设计的调查分析通常不会完全排除具缺失值的观察值，这种分析通过把具缺失值的观察值的权重设为 0 而仍然使用完整的数据库。这对于估计抽样设计中固有的方差是非常有必要的。无论是把观察值完全排除还是把它的权重设为 0，点估计的结果都一样，但方差估计的结果却不相同。把观察值完全排除会低估方差。

如果选项无应答情况比较严重，那么忽略缺失值的做法就会把太多的权重设为 0，这样原来的加权方案就会遭到破坏从而产生偏差，也就不可能再准确反映目标总体。解决这个问题的方法是加大没有缺失值的观察值的权重，以弥补那些被忽略掉的观察值。在实施这种调整时，我们通常假定具缺失值的对象内部并不存在系统的模式，但这个假定也可能无效。比如，如果所有具缺失值的对象都是男性或者都属于某个特定的年龄群体，那么简单地加大那些剩下的观测值的权重就不合适。另一个可行的权重调整办法是利用某些合理的方法填补这些缺失值，但这个方法并不一定比加权调整的方法好。

缺失值的填补不是常规的统计工作。填补缺失值的方法有很多种(Kalton & Kasprszky, 1986; Little & Rubin, 2002)。在选择某种特定的填补方法之前，非常有必要先了解导致缺失值的机制。在某些情况下，我们可以使用简单的方法。比如，对类别变量来说，可以多加一个类别表示缺失值。对连续变量，另一个简单的办法是均值填补，但是这种方法会扭曲变量的分布。为了保持原分布，我们可以使用热卡填补、回归填补，或者多重填补，下一章我们将对填补方法进行简单(并不深入的)说明。如果使用了填补法，方差估计就可能需要某些调整(Korn & Graubard,1999,第 5.5 节)，但这个问题并不在本书的讨论范围之内。

在任何实质性分析之前，很有必要检验一下是否每个 PSU 都有足够的观察值。由于无应答和缺失值的排除，某些 PSU 很可能只包含非常少的观察值，甚至根本没有观察值。一个没有观察值或者只有很少观察值的 PSU 可以与同一

阶层内相邻的 PSU 合并起来。由于合并 PSU，因此只有一个 PSU 的阶层会与相邻的阶层合并起来。但是，合并太多的 PSU 和阶层会破坏原来的抽样设计。随之而来的数据分析也可能出现问题值，因为已经不可能再判断（PSU/阶层）合并后的样本代表的究竟是哪个总体。每个 PSU 所需要的观察值的数量取决于所要进行的分析类型。分析性研究比描述性统计量的估计所需要的观察值数量更大。普遍的法则是这个数量应该足够大，以对某个估计值计算 PSU 内部方差。

为了说明这点，我们来看看 GSS 数据。按阶层和 PSU 进行的未加权列表结果显示，PSU 内部观察值的数量在 8 至 49 之间，大部分大于 13，这意味着 PSU 很可能已经足够大，可用以估计均值和百分比的方差。在分析性研究中，我们可能需要分析按教育程度和性别分类后同意攻击（别人）的成年人的百分比。对于这个分析我们需要检查在特定的分性别教育类别内，是否存在一些没有任何观察值的 PSU。如果对某些分性别教育类别存在许多没有观察值的 PSU，这就会使方差—协方差矩阵的估计（这个估计基于阶层内 PSU 总体的变异情况）出现问题。按 PSU 得出的分性别（2 类）教育（3 类）列表结果显示，84 个 PSU 中有 42 个至少有一个分性别教育单元没有观察值。即使把教育程度合并成两类，我们也需要合并几乎一半的 PSU。因此，我们不应就关于攻击这个问题同时分析性别和教育这两个变量与攻击问题的关系，但可以在不需要合并较多 PSU 和阶层的情况下，单独分析性别或者教育与攻击问题的关系。

我们不能通过在分析范围内选取观察值的方法进行复杂调查数据的子群体分析。虽然个案选择不会改变基本权重，但它可能会破坏基本的抽样设计。比如，选择一个小的种族类别可能会去掉一部分 PSU，剩下的 PSU 中观察值的数量也会显著减少。结果难以从选到的子集中评估基本设计中固有的设计效应。即使基本设计并没有完全被破坏，但从复杂调查样本中选择观察值也可能导致方差的不准确估计，正如上面处理缺失值的方法中所述的一样。准确的方差估计需要在分析中保持完整的数据并对在分析范围外的观察值赋予 0 权重。软件包中可用的子总体分析方法，其子群体分析并不对分析范围内的观测值进行选择。我们将在第 6 章对子总体分析做进一步介绍。

前期分析的第一步是检查主要变量的基本分布。列表可以指出是否需要修正变量的操作定义以及是否需要对特定变量合并其取值类别。基于汇总统计,我们可以知道样本中某些变量的有意思的模式和分布情况。在对变量进行逐一分析之后,我们接下来可以检验变量间是否存在关系,以排除那些并不相关的变量或者与因变量无关的变量。我们可以用标准的、以 SRS 为基础的图形统计方法进行前期检验。但是,由于调查数据中权重的重要作用,任何忽略了权重前期分析的检验都不可能达到预期目标。一种把权重考虑在内的前期分析方法是选择具可操作规模的子样本,其选择概率与权重大小成比例,以利用标识统计图形方法分析这个子样本。第 6 章第一部分将对这个方法进行介绍。

方差估计方法的选择

把设计特征结合到分析中去需要对方差估计方法进行选择。如第 4 章所述,在实际中有三种方差估计方法(BRR, JRR 和泰勒级数近似法)。不同的研究者(Bean, 1975; Frankel, 1971; Kish & Frankel, 1974; Lemeshow & Levy, 1979)对这三种方法进行了经验性评估,Krewski 和 Rao(1981)还对这些方法做了一些理论上的比较。这些评估研究显示,这三种方法中没有哪一种能始终比其他两种更好或者更坏,在多数情况下这些方法的选择可能取决于软件的适用性或者对软件的熟悉情况。少数情况下,这些方法的选择可能取决于需要进行估计的统计量的类型或者使用的抽样设计,就像在配对选择设计中一样。

基于公式的泰勒级数近似法可能是应用最广泛的复杂调查的方差估计方法,因为它在大部分可用的软件中都能找到。由于现实原因,它可能比基于复制的方法(BRR 和 JRR)更好,但正如第 4 章所述,它不适用于中位数或其他百分位数以及非参数统计量。基于复制的方法更普遍,且可应用于这些统计量,但它们需要创建和处理复合样本。基于复制的方法的另一个优点是它提供了一个简单的方法以更加合理地融入无应答和事后分层调整。通过单独对每个复合样本进行加权调整,它可以在方差估计中融入调整的效应。

对小型调查和小范围估计来说,JRR 可能比 BRR 更稳定,因为在 JRR 中每个复合样本都包含了大部分全样本,而 BRR 复合样本只包含一半的样本。但对于百分位数估计,BRR 比 JRR 更加可靠(Kovar et al. , 1988; Rao, Wu & Yue, 1992)。Fay 建议的 BRR 的一种变体(Judkins,1990)被用于稳定方差估计。在这个变体中,复合权重被加大了,被乘以($2-k$) 或 k 而不是 2 或者 0;而且方差估计也被调整了,即把方程 4.4 的右边乘以 $1/(1-k)^2$ (k 的取值在 0 和 1 之间)。Korn 和 Graubard(1999: 35—36)发现,当 $k=0.3$ 时,Fay 的 BRR 会产生与标准 BRR 差不多相同的结果,但当把无应答和事后分层调整结合到复合权重中时,会得到稍微小一点的方差。Fay 的方法可以视为是 JRR 和 BRR 的折中。Judkins(1990)发现,对于百分位数和其他统计量的估计,当 $k=0.3$ 时,Fay 的方法在偏差和稳定性方面比标准 BRR 或者 JRR 都更好。

BRR 的设计目的是用于配对选择设计。当我们从每个阶层中选择 1 个 PSU 时,必须对这些 PSU 进行匹配以创建拟阶层从而使用 BRR 方法。当从每个阶层中选出来的 PSU 的个数大于 2 时,就难以建立配对设计,使用 JRR 或者泰勒级数近似法会更好。从程序上说,泰勒级数近似法是最简单的,而基于复制的方法需要更多的步骤以创建复合权重。

可用的计算资源

过去的 30 年,有许多不同的程序被开发出来用以分析复杂调查数据。早期的程序在政府机构、调查研究组织和大学中因不同的目的而开发出来。其中某些程序进一步开发成针对普通用户的程序包。最初的程序包是为主机计算应用而开发出来的。随着个人电脑计算能力的增强,更多的努力投向了 PC 版本的开发,一些新的软件包也随之出现。虽然其中一些已经不再更新,但还有三种程序包与软件标准的现有状况保持一致,而且便于使用。它们是 SUDAAN, Stata 和 WesVar。

SUDAAN 软件包已经流行了 20 多年,可从北卡罗来纳的三角研究所获取。它有两个版本:单机版和 SAS(统计分析系统)版。后者结合 SAS 数据步特

别便于使用。如 SAS 一样,SUDAAN 的许可证需要每年更新一次。

方差估计的默认方法是泰勒级数近似法,也可选择使用 BRR 或 JRR。它几乎可以处理所有类型的抽样设计,包括多层嵌套和事后分层设计。它具有可用于复杂调查数据分析的最全面的分析特征,但相比另外两个软件包,它的维护需要更多的费用。它支持一系列不同的统计程序,包括 CROSSTAB, DESCRIPT, RATIO, REGRESS, LOGISTIC, LOGLINK(对数线性回归), MULTILOG(多类别和次序 logistic 回归)和 SURVIVAL(Cox 比例风险模型)。过去的这些年,设计者加进了一些新的程序,去掉了一些旧的程序,如用于加权最小二乘建模的 CATAN。SAS 使用者可能会觉得这些程序很容易执行,但对初学者来说,如果没有有经验的用户帮助,可能难以明确该选用哪种设计,也难以解释输出结果。SAS 网站提供了咨询和技术支持。

Stata 是包含了调查分析模板的通用的统计程序包,可从德克萨斯州大学城 Stata 公司获取。它的调查分析部分支持一系列不同的分析性程序,包括 svymean, svytotal, svyprop(比例), svyratio(比率估计), svytab(二维表), svyregress(回归), svylogit(logistic 回归), svymlogit(多类别 logistic 回归), svypois(poisson 回归), svyprobit(probit 模型)以及其他。它利用 PSU 用泰勒级数近似法进行方差估计(最终集群逼近法)。虽然它并不支持复杂的设计如多层嵌套设计,但可用于分析大部分实用的调查设计。其大部分调查分析程序与它的通用(非调查)统计程序类似,这意味着,它的统计分析中的许多普遍特征可轻易与调查分析部分结合起来。它的输出结果也相对比较容易理解,因此新手可能会觉得 Stata 比 SUDAAN 更容易学。

WesVar 程序是由美国马里兰州罗克维尔市的 Westat 公司(Westat, Inc., Rockville, Maryland)研发和推广的。它利用复制的方法计算描述性统计量、线性回归建模和对数线性模型。有 5 种可用的复制方法,包括 JK1(用于无分层设计的 delete-1 jackknife), JK2(用于每层 2 个设计的 jackknife), JKn(用于分层设计的 delete-1 jackknife), BRR 和 Fay(使用 Fay 的方法的 BRR)。它由 Westat 公司研发(Flyer & Mohadjer, 1988),此软件包的旧版本可免费获取。现在这个软件包在市场上可以买到,而且还有一个学生版本。虽然可从其他系统中

导入数据，但这个软件包最初的设计模式是单机版软件包。除了需要全样本的抽样权重，它还要求输入数据文件中的每条记录都包含复合权重。对简单设计来说，它可以在运行任何程序之前先创建复合权重。关于这个软件的文档信息是很充足的，但如果没有有经验用户的帮助，初学者可能会觉得关于创建复合权重的描述有点难以理解。

Cohen(1997)就编程难度、效率、精确度和程序设计能力等方面，评估了这三种软件包的早期版本(SUDAAN 版本 7，Stata 版本 5，WesVarPC 版本 2.02)。评估结果显示，WesVar 程序一贯要求用最少的编程命令得出需要的调查估计值，但它要求准备附加数据以创建推导方差估计所必需的复合权重。Stata 需要更多的命令来得出同样的结果，但它不会在执行命令时产生过分的压力。就计算效率来说，SUDAAN 在产生所需的估计值时始终占优势。

很难说哪一种软件包比其他软件包更值得推荐。在选择软件包时，我们应该在分析需要的背景下，考虑要用到的方差估计方法、维持这个软件的费用以及之前所讲到的这些软件包各自的优点和缺陷。更重要的一点也许是分析者对软件包的熟悉程度。比如对 SAS 用户来说，选择 SUDAAN 是很自然的。Stata 用户也很可能会选择使用 Stata 的调查分析部分。第 6 章对 Stata(版本 8)和 SUDAAN(版本 8.0.1)的使用做了详细说明，这些说明会为软件包的选择提供额外的建议。

随着更多的统计软件包融入复杂调查数据分析程序，计算资源的可用性一直在不断完善。SPSS 13.0 现在给出了一个附加模块用于调查数据分析——SPSS 复杂样本模块(SPSS Complex Samples)。它包含 4 个程序：CSDESCRIPTIVES，CSTABULATE(列联表分析)，CSGLM(回归，ANOVA，ANCOVA)和 CSLOGISTIC。SAS 系统也在最新版本 SAS9.1 中提供了调查数据的分析能力。SURVEYFREQ 程序能产生具相关检验的单向和多维列联表分析。SURVEYLOGISTIC 程序执行logistic 回归，也适用于其他联结函数。目前我们可以利用许多现成可用的统计软件包来进行调查数据分析，而不需要任何特设的软件。

创建复合权重

BRR 方法需要复合权重。这些权重可能包含在数据库中或者在运行分析之前创建。通常需要创建 H(4 的倍数,大于阶层的数量)组复合权重。对于 SUDAAN 来说,复合权重输入在数据之中,WesVar 可以用适当的设置创建复合权重。比如,对于具 6 层且每层含 2 个 PSU 的调查来说,如表 5.1 所示,需要建立 8 组复合权重。这个例子中 BRR 的 SUDAAN 命令列在表格的左边[①]。输

表 5.1　BRR 复合权重的创建和 Jackknife 程序

(A) BRR 的 SUDAAN 命令

```
data brr;
input stratum psu beds aids wt w1-w8;
aids2=aids*2501;
datalines;
1 1  72 20 2  0  0  0  4  0  4  4  4
1 2  87 49 6 12 12 12  0 12  0  0  0
2 1  99 38 2  4  0  0  0  4  0  4  4
2 2  48 23 2  0  4  4  4  0  4  0  0
3 1  99 38 2  4  4  0  0  0  4  0  4
3 2 131 78 4  0  0  8  8  8  0  8  0
4 1  42  7 2  0  4  4  0  0  0  4  4
4 2  38 28 2  4  0  0  4  4  4  0  0
5 1  42 26 2  4  0  4  4  0  0  0  4
5 2  34  9 2  0  4  0  0  4  4  4  0
6 1  39 18 4  0  8  0  8  8  0  0  8
6 2  76 20 2  4  0  4  0  0  4  4  0
;
proc ratio design=brr deff;
weight wt;
repwgt w1-w8;
numer aids2;
denom beds;
run;
```

BRR 的 SUDAAN 输出结果

```
The RATIO Procedure
Variance Estimation Method: BRR
by: Variable, One.
```

变量		One 1
AIDS2/BEDS	Sample Size	12
	Weighted Size	32.00
	Weighted X-Sum	2302.00
	Weighted Y-Sum	2741096.00
	Ratio Est.	1190.75
	SE Ratio	154.15
	DEFF Ratio #4	2.10

(B) 刀切法的 SUDAAN 命令

```
Data jackknife;
input stratum psu beds aids wt;
aids2=aids*2501;
datalines;
1 1  72 20 2
1 2  87 49 6
2 1  99 38 2
2 2  48 23 2
3 1  99 38 2
3 2 131 78 4
4 1  42  7 2
4 2  38 28 2
5 1  42 26 2
5 2  34  9 2
6 1  39 18 4
6 2  76 20 2
;
proc ratio design=jackknife deff;
nest stratum;
weight wt;
numer aids2;
denom beds;
run;
```

刀切法的 SUDAAN 的输出结果

```
The RATIO Procedure
Variance Estimation Method: Delete-1 Jackknife
by: Variable, One.
```

变量		One 1
AIDS2/BEDS	Sample Size	12
	Weighted Size	32.00
	Weighted X-Sum	2302.00
	Weighted Y-Sum	2741096.00
	Ratio Est.	1190.75
	SE Ratio	141.34
	DEFF Ratio #4	1.76

资料来源:Levy and Lemeshow(1999:384).

---

① 原文中为"右边",有误。——译者注

入的数据包括层，PSU，病床的数量，艾滋病人的数量，抽样权重和8组复合权重(w1到w8)。注意复合权重要么等于样本权重的2倍(如果这个单元被选中)，要么等于0(如果这个单元没有被选中)。这些0或2倍的权重根据6行8×8的正交矩阵进行排列，与表4.2类似。艾滋病人总体的比率估计(再次参考注释2)等于目标区域内艾滋病(床位)与床位总数(2501)的比。PROC RATIO和DESIGN = brr指定说明所要的统计量和估计方差的方法，deff要求得出设计效应。因为使用的是BRR设计，所以不需要NEST这个命令(指定说明阶层和PSU)。REPWGT指定复合权重变量。NUMER和DENOM指定说明比率估计的分子和分母。由SUDAAN而得的BRR估计值见表5.1左下部分。估计值等于1191，其标准误差为154.2(设计效应为2.10)。同样的数据，WesVar基于同一正交矩阵不同的行来创建复合权重，其得出的点估计是一样的，而标准误差为155.0。使用不同的复合权重(抽样权重的1或0倍)的BRR得出的标准误差为149.5。Fay的BRR变体使用的复合权重根据某个观察值是否在被选中的单元内而等于抽样权重的$2-k$或者k倍($0 \leqslant k < 1$)。当$k = 0.3$时，标准误差估计为137.3。

Jackknife程序并不需要复合权重。这个程序通过在每个复合样本中删除1个PSU来创建复合样本。JRR的SUDAAN命令和结果见表5.1的右部分。由Jackknife程序估计到的标准误差为141.3，比BRR估计的结果小。由泰勒级数近似法(假定为有放回抽样)得到的标准误差为137.6，比Jackknife估计值稍微小点，而与Fay的BRR估计结果类似。如第4章所述，BRR和JRR假定有放回抽样。如果我们假定无放回抽样(使用有限总体校正法)，这个例子中的标准误差估计值为97.3。

美国国家卫生统计中心(NCHS)的第三次国民健康与营养调查(NHANES III)[4]包含了BRR的复合权重。这些复合权重是为Fay的方法而创建的，$k = 0.3$，且融入抽样不同阶段的无应答和事后分层调整。如Korn和Graubard(1999)提出的，使用Fay的把无应答和事后分层调整融入在内的创建复合权重的办法更可取。但许多调查数据库通常没有这种权重，也缺少适当的创建这种复合权重的信息。

寻找合适的调查数据分析的模型

据说许多统计分析是在没有确立清楚的目标的基础上开展的。在分析数据之前,好好考虑研究问题并建立一个明确的分析计划是很重要的。如前面章节所述,在调查分析中对数据进行前期分析和探索是非常重要的。相比基于设计的分析,这个任务在基于模型的分析中要困难得多。

问题的提出可能会涉及提出问题或者开展合适的背景研究以获取选择某个合适的模型所需要的信息。调查分析者通常没有参与数据收集,要理解数据收集的设计通常很困难。对最初的设计进行提问可能并不充分,但却很有必要弄清楚设计方案如何在实地调查中执行。通常,相关的设计信息既无法在相关文档内找到也无法从数据库中得到。再者,给定可能的样本规模,某些调查设定的目标可能也太过不切实际。所谓的通用调查不可能包括后来使用者关心的所有问题。要建立一个包含所有相关变量的合适的模型是一个挑战。

此外还应该检查已有知识,尤其当类似的数据曾在以前被分析过时。最好不要从头开始拟合模型,相反可以检查一下新的数据是否与先前的(研究)结果一致。遗憾的是,在社会科学和健康科学研究中很难找到使用复制调查数据的基于模型的分析。许多处理基于模型分析的文章关注的是在理想化的条件下分析调查数据的最优程序。比如,大部分公共调查数据库只包含阶层和 PSU,多层或者分层线性模型(Bryk & Raudenbush, 1992; Goldstein & Silver, 1989; Korn & Graubard, 2003)界定额外目标参数的机会有限。复杂调查数据分析的混合线性模型的使用需要进一步的研究,我们希望它能促使调查设计者把设计和分析更紧密地结合起来。

第 6 章 | 调查数据分析的操作

这一章会对调查数据分析做各种说明，重点介绍融入权重和数据结构对分析的影响。我们首先介绍一种对大型复杂调查进行前期分析的策略。我们利用 NHANES III 阶段 II 的数据对各种不同的分析加以说明，包括描述性分析、线性回归分析、列联表分析和 logistic 回归分析。对每种分析，我们都将讨论调查数据要求的一些理论上或实际上的考虑因素。分析中所使用的变量的选择是基于阐释方法的需要，并不是为了说明实质性的发现。最后，我们还将就基于模型的视角进行讨论，因为它与本章列举的分析性例子相关。

预备性分析的策略

抽样权重可能会在复杂调查数据的前期分析中造成破坏，但在数据探索的过程中忽略权重并不是令人满意的解决办法。另一方面，用于调查数据分析的程序并不适用于基本的数据探索。而且图解法并不是专门与复杂调查一起设计出来的。在这一部分，我们会介绍一种把权重考虑在内的前期分析策略。

在发明计算机之前，权重的处理在数据分析中有各种不同的方法。当 IBM 分类机器用于数据列表时，常用的做法是复制数据卡片以匹配权重值从而获取合理的估计值。为了加快大型调查的制表速度，某些调查采用了 PPS 方法(Murthy & Sethi，1965)。意识到分析复杂调查数据的困难性，Hinkins，Oh 和 Scheuren(1994)提倡使用逆抽样设计演算法，这种方法可从现有的复杂调查数据中产生一个简单随机子样本，用户因此可以直接把传统的统计方法应用到这个子样本中。这些方法对调查数据分析已经不再有吸引力了，因为现在已经有

现成可用的调查分析程序。但是，由于前期分析没有必要使用完整的数据库，因此用 PPS 方法选取子样本的原理仍然是为前期分析建立样本的一个好方法。

PPS 子样本可用常规的描述性和图表方法进行分析，因为权重已经反映在子样本的选择之中。比如，散点图是前期数据探索的其中一种最重要的图表法。一种把权重结合到散点图中的方法是使用代表权重大小的气泡。Korn 和 Graubard(1998)检验了各种不同的绘制双变量数据散点图的方法，说明了使用 PPS 子样本的优点。实际上，他们发现抽样散点图比气泡散点图更可取。

为了前期分析，我们从 NHANES III 阶段 II(1991—1994)(参见注释[4])的成年人数据(共 9920 个成年人)中产生了一个样本规模为 1000 的 PPS 样本。

表 6.1　某些选定特征的子样本和全样本估计结果：美国 NHANES III 阶段 II 调查成年人人口

	平均年龄	维他命的使用	西班牙人口	SBP[a]	样本 BMI[b] 和 SBP 的相关性
总样本(n=9920)[c]					
未加权	46.9 岁	38.4%	26.1%	125.9 mmHg	0.153
加权后	43.6	42.9	5.4	122.3	0.243
PPS 子样本(n=1000)					
未加权	42.9	43.0	5.9	122.2	0.235

注：a. 收缩压；
b. 体重指数；
c. 大于或等于 17 岁的成人。

我们首先按阶层和 PSU 对总样本进行分类，然后在累积相对权重的取值范围内用 9.92 的跳跃间距系统地选取 PPS 子样本。按阶层和 PSU 进行的分类整理从根本上保留了原有抽样设计的完整性。

表 6.1 显示了可用传统统计软件包进行分析的 PPS 子样本的实用性。我们选择了几个最受权重影响的变量。由于我们对老年人和少数种族进行了过取样，平均年龄和西班牙人口比例的加权后估计值与未加权估计值不一样。此外，权重还使维他命使用情况和收缩压的估计结果不一样，因为它们受到过抽样类别的严重影响。虽然没有经过加权，但子样本中的估计值却与加权后的总样本中的估计值非常接近，显示了前期分析中 PPS 子样本的有用性。在体重指

数与收缩压的关联度这方面也存在类似的结果。

PPS 子样本在并不正式考虑权重而对数据进行探索时非常有用，对还处于入门阶段的学生尤其有用。它特别适合通过如散点图、对比箱形图和中位数标示图等图形方法探索数据。其真正的优势在于重新抽样过的数据近似地代表了总体，可以在忽略权重的情况下进行分析。从这种重新抽样过的数据中得到的点估计几乎与完整数据库中的加权后估计值相等。从这种重新抽样过的数据中发现的有趣的趋势很可能会被更完整的（通过 Stata 或者 SUDAAN 进行的）分析所证实，虽然标准误差可能会有所不同。

描述性分析

我们使用 NHANES III 阶段 II 中的成人样本（17 岁或以上）进行描述性分析。这个样本包含 9920 个观察值，这些观察值处于 23 个拟阶层中，每个阶层含 2 个拟初级抽样单位。拟阶层和初级抽样单位的标识包括在我们的工作数据文件之中。数据文件中扩展权重被转化为相对权重。为了检查在不同的 PSU 内其观察值的分布是否存在问题，我们制作了未加权列表。结果显示，每个 PSU 中包含的观察值的数量在 82 至 286 之间。这些 PSU 的样本规模对进一步分析已经足够。

我们决定分析体重指数、年龄、种族、贫困指数、收缩压、维他命使用情况和抽烟状况。体重指数等于体重（公斤）除以高度（米）的平方。年龄用年测量，教育用受教育年限测量，贫困指数等于家庭收入与贫困程度的比，收缩压用 mmHg 测量。此外，我们还选择了以下几个二分类变量：Black（1＝黑人；0＝非黑人），Hispanic（1＝西班牙裔；0＝非西班牙裔），使用维他命（1＝是；0＝否），抽烟状况（1＝曾经抽过；0＝从未抽过）。

对分析中选中的变量我们填补了其缺失值以说明调查数据分析的步骤。在以往的研究中已有各种不同的填补办法用以弥补缺失的调查数据（Brick & Kalton，1996；Heitjan，1997；Horton & Lipsitz，2001；Kalton & Kasprszky，1986；Little & Rubin，2002；Nielsen，2003；Zhang，2003）。也有几种软件包

可供利用(如 SAS/STAT 中的 proc mi 和 proc mianalyze; SOLAS; MICE; S-Plus Missing Data Library)。有很多方法可把这些软件包应用到某个特定的数据库中。合适的应用方法和(处理)过程的选择最终取决于缺失值的数量、产生缺失值的机制(是否可以忽略)以及缺失值的模式(单调的还是一般的)。复杂统计方法的使用很吸引人,但却可能弊大于利。更好的方法是根据具体的实例(Kalton & Kasprszky, 1986; Korn & Graubard, 1999 第 4.7 节和第 9 章),而不是技术性的操作指引来选择合适的方法。关于这方面详细的讨论不在本章讨论之列。以下简单的介绍仅供解说之用。

在我们的数据中,年龄和种族没有缺失值。我们先填补那些缺失值最少的变量值。维他命使用情况和抽烟状况的缺失值少于 10 个,教育和高度的缺失值大约为 1%。我们使用热卡填补[5](就近填补)的方法对这 4 个变量填补缺失值——按性别在 5 年一组的年龄类别内以与样本权重成比例的概率随机选取捐赠观察值。当某个观察值内还有更多的变量具缺失值时,我们也使用相同的观察值对缺失值进行填补。我们使用回归填补法来填补高度(3.7%缺失;基于权重、年龄、性别和种族则 2.8%缺失;基于年龄、性别和种族则 0.9%缺失)、收缩压(2.5%缺失,基于高度、权重、年龄、性别和种族)、贫困指数(10%缺失,基于家庭规模、教育和种族)等的缺失值。填补的贫困指数值约 0.5%为负值,这些值被设成 0.001(数据中最小的贫困指数值)。附带说明一下,我们可以把其他人体测量的方法放到回归填补法之中,但我们的演示仅基于这个分析所选择的变量。最后,体重指数值(5.5%缺失)以更新后的权重和高度信息为基础重新计算而得。

为了说明抽样权重和设计效应会导致不同的结果,我们采用三种不同的方法进行分析:(1)未加权的,忽略数据结构;(2)加权后的,忽略数据结构;(3)调查分析,结合权重和抽样特征。第一种方法假定简单随机抽样,第二种考虑了权重但忽略了设计效应,第三种则为给定的抽样设计提供合适的分析。

首先,我们在填补和未填补缺失值这两种情况下,检验了加权后均值和比例以及它们的标准误差。填补的结果对点估计并没什么重大影响,用第三种分析法估计到的标准误差稍微变小了一点。没有填补缺失值时,加权后的贫困指数均值

为 3.198(标准误差 s.e. = 0.114)，填补缺失值后结果为 3.168(s.e. = 0.108)。对于体重指数，没有填补缺失值时，加权后的均值为 25.948(标准误差 s.e. = 0.122)，填补缺失值后结果为 25.940(s.e. = 0.118)。其他变量，点估计值及其标准误差在这两种情况下得到的结果到小数点第三位相同，因为其缺失值很少。

估计到的描述性统计值(使用了填补值之后)列于表 6.2。计算使用的是 Stata。表格顶部未加权的统计值是由非调查分析命令 summarize(用于点估计)和 ci(用于标准误差)得出的。表格顶部加权后的估计值(第二种方法)是通过相同的非调查分析命令加上[w=wgt]得到的。第三种分析(把权重和设计特征融入进来)的结果列于表格底部。它使用 svyset[pweight=wgt]，strata(stra)和 psu(psu)设定复杂调查特征，以及 svymean 对具体的变量估计均值和百分比。

表 6.2 中显示的统计值为连续变量的均值估计值、二分类变量的百分比和标准误差估计。有几个变量其加权后和未加权的均值/百分比之间存在细微差别，但对某些变量来说，这种差别非常大。加权后黑人的比例比未加权时的比例小 60%以上，加权后西班牙裔的比例比未加权时的小将近 80%，反映了对这两个少数种族的过抽样。加权后的平均年龄约比未加权的平均年龄小 3.5 岁，原因在于对年老者进行了过抽样。另一方面，对贫困指数和受教育年限来说，加权后的均值比未加权的均值大很多，意味着被过抽样的少数种族集中于收入和教育分布中的底端。加权后的维他命使用情况的估计值也在某种程度上比未加权时大一些。这可能反映了少数种族较少使用维他命。

表格底部显示了(同时反映了权重和设计特征的)调查估计值。虽然估计到的均值和百分比与表格顶部加权后的结果完全相同，但对所有的变量来说，其标准误差都显著增大了。这种差别反映在表格中的设计效应中(表格底部标准误差与表格顶部加权后统计值的标准误差之比的平方)。贫困指数、教育和年龄的巨大的设计效应部分反映了这些特征的居住同质性。这两个社会经济特征变量(贫困指数和教育)和年龄的设计效应比黑人和西班牙裔这两个变量的设计效应要大。而 1976—1980 年的 NHANES II 中的结果却相反(数据用于本书第一版中)，意味着相比种族身份，现在在社会经济身份这方面居住区域更具同质性了。

表 6.2 NHANES III 阶段 II 中 17 岁及以上成人的回归分析中所用变量的描述性统计(用 Stata 进行的分析)

(A) 加权和未加权的描述性统计，忽略设计因素

Variable	Unweighted Analysis Mean	Unweighted Analysis Std. Err.	Weighted Analysis Mean	Weighted Analysis Std. Err.	Min	Max
bmi	26.4465	.05392	25.9402	.05428	10.98	73.16
age	46.9005	.20557	43.5572	.17865	17	90
black	.2982	.00459	.1124	.00317	0	1
hispanic	.2614	.00441	.0543	.00228	0	1
pir	2.3698	.01878	3.1680	.02086	0	11.89
educat	10.8590	.03876	12.3068	.03162	0	17
sbp	125.8530	.20883	122.2634	.18397	81	244
vituse	.3844	.00488	.4295	.00497	0	1
smoker	.4624	.00501	.5114	.00502	0	1

(B) 利用权重和设计因素的调查分析

```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svymean bmi age black hispanic pir educat sbp vituse smoker

Survey mean estimation
pweight:  wgt                          Number of obs(*)  =     9920
Strata:   stra                         Number of strata  =       23
PSU:      psu                          Number of PSUs    =       46
                                       Population size   =  9920.06
```

Mean	Estimate	Std. Err.	[95% Conf.	Interval]	Deff
bmi	25.9402	.11772	25.6946	26.2013	4.9903
age	43.5572	.57353	42.3708	44.7436	10.3067
black	.1124	.00973	.0923	.1326	9.4165
hispanic	.0543	.00708	.0397	.0690	9.6814
pir	3.1680	.10779	2.9622	3.4328	25.6331
educat	12.3068	.12312	12.0565	12.5671	15.0083
sbp	122.2634	.38543	121.4010	122.980	4.1995
vituse	.4295	.01215	.4043	.4546	5.9847
smoker	.5114	.01155	.4874	.5352	5.2829

注：* 有些变量有缺失值。

表格底部还显示了均值和百分比的 95%置信区间。置信限度所用的 t 值并不是我们熟悉的可从 $n = 9920$ 样本中期望得到的 1.96(相对权重的和)。其原因在于，在多阶段集群抽样设计中，自由度是基于 PSU 和阶层的数量，而不是像 SRS 中那样基于样本规模。需要特别强调的是，复杂调查中自由度等于选中的 PSU 的数量减去所使用的阶层的数量。在我们的例子中，自由度为 23(= 46 − 23)，$t_{23,0.975} = 2.0687$，这个 t 值用于表 6.2 中的所有置信区间。在某些情况下，自由度可能会与从上述一般方法中得到的稍微不同(Korn & Graubard, 1999，第 5.2 节)。

在表 6.3 中，我们演示了进行子群体分析的例子。正如前一章所提到的，任何使用复杂调查数据进行的子群体分析都应该使用完整的数据库，而不能在分析范围内任意选取数据。在 Stata 中有两种合适的子群体分析的方法：使用 by 或者 subpop。表 6.3 显示了对黑人这个变量进行子群体分析的例子。在表格顶部，我们用 by 对非黑人和黑人分别估计了 BMI 均值。黑人的 BMI 均值比非黑人的大。虽然在非黑人中 BMI 的设计效应(5.5)与总体的设计效应(表 6.2 中的 5.0)相似，但在黑人中却只有 1.1。

Stata 还可用于检验参数的线性组合。两个子群体之均值是否相等可以用 lincom 命令进行检验(如[bmi]1－[bmi]0，检验关于当 black = 1 和当 black = 0[①] 时各自的总体 BMI 均值之间的差别的假设)，基于 t 检验这两个均值之间的差异在统计上是显著的。

设计效应是 1.46，表示这个检验的 t 值降低了约 20% 以抵消抽样设计特征。

另一种方法，subpop 可用于估计黑人的 BMI 均值，如表格底部所示。这种方法使用了完整的数据库，把不在分析范围内的观察值的权重设为 0。均值、标准误差和设计效应与表格顶部用 by 命令得到的黑人的结果相同。

下面，我们通过明确(分析)范围(if black = 1)把黑人选择出来，从而估计 BMI 均值。因为在某些 PSU 中并没有黑人，所以这个方法并不奏效。按阶层和 PSU 的列表显示，在第 13 个和第 15 个阶层内只有 1 个 PSU。当这两个阶层与邻近的阶层合并起来时，Stata 分析产生了结果。虽然点估计与前面的相同，但标准误差和设计效应却不同。普遍的法则是，调查数据的子群体分析应该避免筛选出子数据库，这与 SRS 数据的分析不同。

除了 svymean 命令外，Stata 还支持以下几种描述性分析：svytotal(用于总体估计)，svyratio(用于比率估计)，svyprop(用于百分比估计)。在 SUDAAN 中，这些描述性统计量可以用 DESCRIPT 程序进行估计，次范围分析则可通过 SUBPOPN 命令实现。

---

① 原文为 nonblack = 0，有误。——译者注

表 6.3 NHANES III 阶段 II 中 17 岁及以上黑人和非黑人平均体重指数的比较 ($n = 9920$)：用 Stata 进行的分析

(A)
```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svymean bmi, by (black)

Survey mean estimation
pweight:   wgt                                  Number of obs    =       9920
Strata:    stra                                 Number of strata =         23
PSU:       psu                                  Number of PSUs   =         46
                                                Population size  =    9920.06
------------------------------------------------------------------------------
Mean   Subpop. |  Estimate    Std. Err.   [95% Conf. Interval]       Deff
---------------+--------------------------------------------------------------
bmi    black==0 |  25.7738     .12925     25.5064    26.0412        5.512
       black==1 |  27.2536     .17823     26.8849    27.6223        1.071
------------------------------------------------------------------------------
```

(B)
```
. lincom [bmi]1-[bmi]0, deff

 ( 1) - [bmi]0 + [bmi]1 = 0.0
------------------------------------------------------------------------------
       Mean |  Estimate   Std. Err.     t     P>|t|   [95% Conf. Interval]   Deff
------------+-----------------------------------------------------------------
        (1) |   1.4799    .21867      6.77    0.000     1.0275    1.9322    1.462
------------------------------------------------------------------------------
```

(C)
```
. svymean bmi, subpop(black)

Survey mean estimation
pweight:   wgt                                  Number of obs    =       9920
Strata:    stra                                 Number of strata =         23
PSU:       psu                                  Number of PSUs   =         46
Subpop.:   black==1                             Population size  =    9920.06
------------------------------------------------------------------------------
    Mean |  Estimate    Std. Err.   [95% Conf. Interval]         Deff
---------+--------------------------------------------------------------------
     bmi |  27.2536     .17823      26.8849    27.6223           1.071
------------------------------------------------------------------------------
```

(D)
```
. svymean bmi if black==1
stratum with only one PSU detected
```

(E)
```
. replace stra=14 if stra==13
(479 real changes made)
. replace stra=16 if stra==15
(485 real changes made)

. svymean bmi if black==1

Survey mean estimation
pweight:   wgt                                  Number of obs    =       2958
Strata:    stra                                 Number of strata =         21
PSU:       psu                                  Number of PSUs   =         42
                                                Population size  =   1115.244
------------------------------------------------------------------------------
    Mean |   Estimate    Std. Err.    [95% Conf. Interval]         Deff
---------+--------------------------------------------------------------------
     bmi |   27.2536     .17645       26.8867    27.6206           2.782
```

线性回归分析

回归分析和ANOVA都检验一个连续因变量和一系列自变量之间的线性相关。为了检验假设，假定因变量服从正态分布。下面这个方程表示了这些方法所考虑的相关方式。对于 $i = 1, 2, \cdots, n$，

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i \qquad [6.1]$$

这是一个线性模型，因为因变量 Y_i 由 β_j 的线性组合加上 ε_i 表示。β_j 是方程中自变量 X_j 的系数，ε_i 是模型中的随机误差项——假定服从均值为0方差为常数的正态分布且相互独立。

在回归分析中，自变量要么是连续的，要么是离散的，而 β_j 就是这些自变量对应的系数。在ANOVA中，自变量 X_j 是指示变量(在效果编码下，自变量的每个类别都分别有一个编码为1或者0的指示变量)，以显示哪种(因素的)效应被加入到模型中，而 β_j 就是这些效应。

当数据来自SRS时，普通最小二乘估计用于获得回归系数的估计值或线性模型中的效应。但在处理从复杂抽样中得到的数据时，就需要对这种方法进行一些修改。现在我们用的数据包含个体观察值加上抽样权重和设计描述变量。正如第3章所讨论的，复杂抽样中个体被选中的概率常常不同。此外，由于抽样设计方面的原因，在复杂调查中随机误差项通常不再相互独立。正因为这些与SRS偏离的情况，模型参数的OLS估计及其方差估计是有偏差的。因此，置信区间和假设检验就可能存在误导性。

很多学者已经讨论过这些问题(Binder, 1983; Fuller, 1975; Holt, Smith & Winter, 1980; Konijn, 1962; Nathan & Holt, 1980; Pfeffermann & Nathan, 1981; Shah, Holt & Folsom, 1977)，他们并没有统一使用某种单一的分析方法，但都认为OLS这种估计方法可能(在某些情况下)不合适。我们在此并不详细回顾所有前人的相关研究，而只重点关注一种可用于最广泛的情况之中的、有现成软件可供利用且已被广泛应用的方法。这种模型参数估计方法是

设计加权最小二乘法，SUDAAN、Stata 和其他复杂调查数据分析软件都支持这种方法的使用。

DWLS 方法中的权重是第 3 章所讨论的抽样权重。DWLS 与不等方差的加权最小二乘法稍微不同，它是从假定的协方差结构中得到权重的（见 Lohr，1999，第 12 章）。考虑到由抽样设计和其他对权重的调整所带来的复杂性，第 4 章所讨论的其中一种方法可用于在模型参数估计值的方差—协方差矩阵估计中。由于这些模型使用的是 PSU 总体而不是个体值作为方差计算的基础，因此这个设计的自由度等于 PSU 的数量减去阶层的数量而不是样本规模。与误差平方和相关的自由度则等于 PSU 的数量减去阶层的数量再减去模型中公式项的数量。

表 6.4 显示了用前面提到的三种分析方法对 BMI 进行多元回归分析的结果。自变量与前面的描述性分析所使用的相同。此外，我们还包括了年龄平方，以控制可能存在的年龄对 BMI 的非线性效应。简单起见，这个例子中没有包含交互项，因为包含交互项除了会加深多重共线性问题外，还会增加 R^2。我们在这个分析中使用了缺失值的填补值。回归系数几乎与没有使用填补值的分析结果相同，系数的标准误差也类似。

表格的顶部显示了不考虑设计特征时未加权和加权后的分析结果。regress 命令同时用于未加权和加权后分析，加权后分析中权重由（w = wgt）命令确定。首先，我们重点来看令人失望的很低的 R^2 值，在未加权分析中是 0.12，在加权后分析中为 0.14。这表示大部分 BMI 的变异情况无法由模型解释。模型还包括了其他重要的变量。也许令人满意的预测 BMI 的模型设定无法用 NHANES III 数据实现。

未加权和加权后分析两者都显示年龄与 BMI 正相关，而年龄平方则与 BMI 负相关。这表示年龄的作用是曲线的，正如我们期望的，年龄越大趋势越弱。贫困指数和教育与 BMI 负相关。至于二元变量的回归系数，黑人和西班牙裔这两个变量的系数都是正的，表示这两个少数种族群体分别比他们的对立群体（非黑人、非西班牙裔）具有更大的 BMI。收缩压与 BMI 正相关，而使用维他命的人（他们可能更加关注自己的健康）则比不使用的人具有更低的 BMI。那些

曾经抽烟的人与没有抽过烟的人相比，二者之间的 BMI 值差距不超过 0.5 个点。

表 6.4 NHANES III 阶段 II 中成人体重指数的多元回归模型 ($n = 9920$)：用 Stata 进行的分析

(A) 不加权和加权分析，忽略设计因素

```
                Unweighted analysis                         Weighted analysis
         ------------------------------------------   -------------------------------------
    Source |       SS         df         MS            |       SS         df         MS
-----------+--------------------------------          -+--------------------------------
     Model |  33934.57         9   3770.48             |  37811.46         9   4201.27
  Residual | 252106.39      9910     25.44             | 236212.35      9910     23.84
-----------+--------------------------------          -+--------------------------------
     Total | 286040.68      9919     28.84             | 274023.81      9919     27.63
---------------------------------------------          --------------------------------
             F(  9,   9910) =  148.21                    F(  9,   9910) =  176.26
             Prob > F       =  0.0000                    Prob > F       =  0.0000
             R-squared      =  0.1186                    R-squared      =  0.1380
             Adj R-squared  =  0.1178                    Adj R-squared  =  0.1372
             Root MSE       =  5.0438                    Root MSE       =  4.8822
------------------------------------------------    ------------------------------------
       bmi |    Coef.   Std. Err.      t    P>|t|      Coef.   Std. Err.      t    P>|t|
-----------+------------------------------------    ------------------------------------
       age |   .38422   .01462     26.27   0.000      .39778   .01528     26.03   0.000
     agesq |  -.00391   .00014    -27.61   0.000     -.00421   .00016    -27.06   0.000
     black |  1.15938   .13178      8.80   0.000      .96291   .16108      5.98   0.000
  hispanic |   .70375   .14604      4.82   0.000      .64825   .22761      2.85   0.004
       pir |  -.14829   .03271     -4.53   0.000     -.12751   .02758     -4.62   0.000
    educat |  -.00913   .01680     -0.54   0.587     -.11120   .01865     -5.96   0.000
       sbp |   .05066   .00313     16.18   0.000      .07892   .00338     23.35   0.000
    vituse |  -.72097   .10752     -6.71   0.000     -.64256   .10176     -6.31   0.000
    smoker |  -.47851   .10456     -4.58   0.000     -.34981   .10033     -3.49   0.001
     _cons | 12.70443   .49020     25.92   0.000    10.36452   .52213     19.85   0.000
------------------------------------------------    ------------------------------------
```

(B) 利用数据设计因素的分析

```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svyregress bmi age agesq black hispanic pir educat sbp vituse smoker, deff

Survey linear regression
pweight:   wgt                                       Number of obs      =      9920
Strata:    stra                                      Number of strata   =        23
PSU:       psu                                       Number of PSUs     =        46
                                                     Population size    =   9920.06
                                                     F(    9,      15)  =     71.84
                                                     Prob > F           =    0.0000
                                                     R-squared          =    0.1380
-------------------------------------------------------------------------------------
       bmi |    Coef.    Std. Err.      t     P>|t|     [95% Conf. Interval]     Deff
-----------+-------------------------------------------------------------------------
       age |   .39778    .02110     18.85    0.000      .35412      .44143     2.0539
     agesq |  -.00421    .00023    -18.02    0.000     -.00469     -.00373     2.3647
     black |   .96291    .22418      4.30    0.000      .49916     1.42666     1.5778
  hispanic |   .64825    .20430      3.17    0.004      .22562     1.07087      .8897
       pir |  -.12751    .05624     -2.27    0.033     -.24855     -.01117     4.5323
    educat |  -.11203    .02703     -4.11    0.000     -.16712     -.05529     2.1457
       sbp |   .07892    .00514     15.35    0.000      .06828      .08956     1.8798
    vituse |  -.64256    .17793     -3.61    0.001    -1.01063     -.27449     3.0546
    smoker |  -.34982    .20405     -1.71    0.100     -.77192      .07229     4.0343
     _cons | 10.36452    .80124     12.94    0.000     8.70704    12.02201     2.3041
-------------------------------------------------------------------------------------
```

未加权和加权后分析间存在着小差别。虽然在未加权分析中，教育的作用很小(beta 系数 =－0.009)，但在加权后分析中其绝对值则变大了很多(beta 系数 =－0.111)。如果前期分析并没有使用抽样权重，我们就可能忽略了教育作为一个重要预测因素的作用。这个例子清楚显示了(本章开头所讨论的)利用 PPS 子样本进行前期分析的优势。抽烟状态的负系数值稍微减弱了，意味着相比其他人，抽烟对 BMI 的负面影响在被过抽样的人群中更加严重。这里再次说明了抽样权重的重要性。这个分析还指出了在前期分析中使用 PPS 子样本而不是未加权分析的优势。

把权重和设计特征同时考虑进来的分析结果列在表格底部。这个分析使用的是 svyregress 命令。回归系数的估计值和 R^2 与加权后分析中的结果一样，因为在这个估计中我们使用的是相同的公式。但是，系数的标准误差和 t 统计值则与加权后分析中的非常不一样。估计到的回归系数的设计效应在 0.89 (Hispanics)和 4.53(pir，贫困与收入比)之间。我们再次看到相对 SRS 设计，复杂调查设计可能会对某些变量带来更大的方差，但并不是对所有的变量都这样。在这个特殊的例子中，前期分析中得到的一般的分析结果在最终的分析中都得到了支持，虽然对除了一个变量以外的所有其他变量来说回归系数的标准误差都增大了。

比较表 6.2 和 6.4 中的设计效应，我们发现，回归系数的设计效应比均值和百分比的稍微小点。所以，把从均值和总体中估计到的设计效应应用到回归系数中(当数据中不包含集群信息时)会导致太保守的结论。如果回归模型控制了某些集群对集群的变异性，那么就可能在回归分析中得到更小的设计效应。比如，如果处于相同集群中的人具有类似 BMI 的部分原因是因为他们具有类似的年龄和教育程度，那么在回归模型中按年龄和教育程度进行调整就可能解释掉一部分集群对集群的变异性。这样集群效应对模型的残差的影响就可能更小。

回归分析也可以用 SUDAAN 的 REGRESS 程序执行，过程如下：

```
PROC REGRESS DESIGN = wr;
NEST stra psu;
```

```
WEIGHT wgt;
MODEL = bmi age agesq black hispanic pir educat sbp vituse smoker;
RUN;
```

列联表分析

分析两个离散变量间相关性的最简单的方法是二维表格。如果数据来自SRS,我们可以用 Pearson 卡方统计值检验关于独立的零假设。对基于复杂调查数据的二维列表分析,需要对这个检验方法进行修改以考虑调查设计的影响。已有多种不同的检验统计值被提出来。Koch,Freeman 和 Freeman(1975)提议使用 Wald 统计值[6],且已被广泛使用。Wald 统计值通常被转化成 F 统计值以决定 p 值。在 F 统计值中,分子中的自由度与表格的维度联系在一起,分母中的自由度反映的是调查设计。后来,Rao 和 Scott(1984)提出了对数似然统计值的修正方法,使用具非整数自由度的 F 统计值。基于一个模拟研究(Sribney, 1998), Stata 把 Rao-Scott 修正统计值设为默认程序,但 Wald 卡方和对数线性 Wald 统计值也是可用的选项。另一方面,SUDAAN 在它的 CROSSTAB 程序中用 Wald 统计值。在多数情况下,这些统计值得到的结论是相同的。

表 6.5 列出了用 Stata 进行的二维列表分析情况。这个分析列出了维他命使用与编码为 3 个类别的教育年限(1=少于 12 年,2=12 年,3=多于 12 年)之间的相关情况。在(A)部分,普通的卡方分析忽略了权重和数据结构的影响。教育与维他命使用之间在统计上存在着显著相关,教育程度越高的人越倾向于使用维他命。维他命使用者的百分比处于 32%(在最低的教育水平中)到 49%(在最高的教育水平中)之间。(B)部分列出了用相同的数据但把调查设计考虑进来时的分析结果。按教育程度划分,加权后维他命使用者的比例其变化情况比未加权时稍微大了一点,处于 33%(第 1 级教育程度)到 52%(第 3 级教育程度)之间。注意当使用了 ci 命令时,Stata 会对单元百分比计算置信区间。

表 6.5 NHANES III 阶段 II 中按教育程度划分的美国成人的维他命使用情况比较 ($n = 9920$)：用 Stata 进行的分析

(A)
```
. tab vituse edu, column chi
-----------------------------------------------------------
           |                 edu
    vituse |         1          2          3 |     Total
-----------+---------------------------------+-----------
         0 |      2840       1895       1372 |      6107
           |     68.43      61.89      50.66 |     61.56
-----------+---------------------------------+-----------
         1 |      1310       1167       1336 |      3813
           |     31.57      38.11      49.34 |     38.44
-----------+---------------------------------+-----------
     Total |      4150       3062       2708 |      9920
           |    100.00     100.00     100.00 |    100.00
          Pearson chi2(2) = 218.8510    Pr = 0.000
-----------------------------------------------------------
```

(B)
```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svytab vituse edu, column ci pearson wald

pweight:  wgt                                  Number of obs      =      9920
Strata:   stra                                 Number of strata   =        23
PSU:      psu                                  Number of PSUs     =        46
                                               Population size    =   9920.06
--------------------------------------------------------------------------
          |                            edu
   vituse |              1              2              3          Total
----------+---------------------------------------------------------------
        0 |          .6659          .6018          .4834          .5705
          | [.6307,.6993]  [.5646,.6379]  [.4432,.5237]  [.5452,.5955]
        1 |          .3341          .3982          .5166          .4295
          | [.3007,.3693]  [.3621,.4354]  [.4763,.5568]  [.4045,.4548]
    Total |              1              1              1              1
--------------------------------------------------------------------------
  Key:  column proportions
        [95% confidence intervals for column proportions]
  Pearson:
    Uncorrected   chi2(2)          =  234.0988
    Design-based  F(1.63, 37.46)   =   30.2841     P = 0.0000
  Wald (Pearson):
    Unadjusted    chi2(2)          =   51.9947
    Adjusted      F(2, 22)         =   24.8670     P = 0.0000
--------------------------------------------------------------------------
```

(C)
```
. svyset [pweight=wgt], strata(stra) psu(psu)
. svytab vituse edu, subpop(hispanic) column ci wald

pweight:  wgt                                  Number of obs      =      9920
Strata:   stra                                 Number of strata   =        23
PSU:      psu                                  Number of PSUs     =        46
                                               Population size    =   9920.06
Subpop.:  hispanic==1                          Subpop. no. of obs =      2593
                                               Subpop. size       =   539.043
--------------------------------------------------------------------------
          |                            edu
   vituse |              1              2              3          Total
----------+---------------------------------------------------------------
        0 |          .7382          .6728          .5593          .6915
          | [.6928,.7791]  [.6309,.7122]  [.4852,.6309]  [.6509,.7293]
        1 |          .2618          .3272          .4407          .3085
          | [.2209,.3072]  [.2878,.3691]  [.3691,.5148]  [.2707,.3491]
    Total |              1              1              1              1
--------------------------------------------------------------------------
  Key:  column proportions
        [95% confidence intervals for column proportions]
  Wald (Pearson):
    Unadjusted    chi2(2)          =   47.1625
    Adjusted      F(2, 22)         =   22.5560     P = 0.0000
--------------------------------------------------------------------------
```

这个分析要求同时报告 Pearson 和 Wald 卡方统计值。基于加权后频数的未经修正的 Pearson 卡方比(A)中的卡方值稍微大点，反映出加权后百分比的变化情况要大些。但是，基于未经修正的 Pearson 卡方统计值，我们无法估计出能反映复杂设计的合适的 p 值。合适的 p 值可以从基于设计的自由度为 1.63 和 37.46 的 F 统计值 30.28 中评估出来，它以由 Rao-Scott 修正产生的检验程序为基础。未经调整的 Wald 卡方检验统计值为 51.99，但 p 值的产生必须以调整后的 F 统计值为基础。两个 F 统计值中的分母自由度都反映了抽样设计中 PSU 和阶层的数量。调整后的 F 统计值只比 Rao-Scott 的 F 统计值稍微小了一点。这些检验统计值中的任意一种都会得出相同的结论。

在(C)部分，我们对西班牙裔人口进行了子群体分析。在这个分析中我们使用了完整的数据文件。分析基于 2593 个观察值，但当考虑了抽样权重之后，只代表 539 个人。西班牙裔中维他命使用者的比例(31%)显著低于总体人口中维他命使用者的比例(43%)。正如调整后的 F 统计值所示，西班牙裔人口中，教育与维他命使用之间在统计上仍然存在显著相关。

现在我们来看看三维列表。利用 NHANES III 阶段 II 中的成人样本数据，我们检验按教育程度划分的维他命使用的性别差异情况。这是一个 2×2×3 的表格，我们可以在每个教育程度中都进行一个二维列表分析。表 6.6 列出了用 SAS 和 SUDAAN 进行的 3 个 2×2 列表分析结果。表格顶部显示的是忽略了调查设计的分析结果。在最低的教育水平中，男性维他命使用者的百分比比女性的低，卡方统计值显示这个差别在统计上是显著的。另一个检验 2×2 列表中的相关性的方法是计算比数比。

在这个表格中，男性维他命使用情况的比数为 $0.2634/(1-0.2634)=0.358$，女性的则为 $0.3617/(1-0.3617)=0.567$。男性比数与女性比数的比为 $0.358/0.567=0.63$，意味着男性使用维他命的比数是女性使用维他命比数的 63%。95%的置信区间内不包含 1，表示这种差别在统计上是显著的。在三个不同程度的教育水平间，比数比结果是一致的。由于这些比数比是一致的，因此我们可以把这三个不同教育程度的 2×2 列表合起来。然后就可计算 Cochran-Mantel-Haenszel(CMH)卡方($df=1$)和 CMH 普通比。按教育调整后的

比数比为 0.64，其 95%置信区间不包含 1。

表 6.6 NHANES III 阶段 II 中按教育程度划分的美国成人的维他命使用性别差异分析 ($n = 9920$)：用 SAS 和 SUDAAN 进行的分析

(A) **用 SAS 做的未加权的分析**

```
proc freq;
tables edu*sex*vituse / nopercent nocol chisq measures cmh;
run;
          [Output summarized below]
```

Level of education:	Less than H.S.		H.S. graduate		Some college	
Vitamin use status:	(n)	User	(n)	User	(n)	User
Gender- Male:	(1944)	26.34%	(1197)	31.91%	(1208)	43.54%
Female:	(2206)	36.17	(1865)	42.09	(1500)	54.00
Chi-square:	46.29		32.02		29.27	
P-value:	<.0001		<.0001		<.0001	
Odds ratio:	0.63		0.64		0.66	
95% CI:	(0.56, 0.72)		(0.56, 0.75)		(0.56, 0.76)	

CMH chi-square: 107.26 (p<.0001)
CMH common odds ratio: 0.64 95%CI: (0.59, 0.70)

(B) **用 SUDANN 做的调查分析**

```
proc crosstab design=wr;
nest stra psu;
weight wgt;
subgroup edu sex vituse;
levels    3   2    2;
tables edu*sex*vituse;
print nsum wsum rowper cor upcor lowcor chisq chisqp cmh cmhpval;
run;
          [Output summarized below]
```

Level of education:	Less than H.S.		H.S. graduate		Some college	
Vitamin use status:	(n)*	User	(n)*	User	(n)*	User
Gender- Male:	(1274.9)	28.04%	(1432.2)	32.66%	(2031.1)	45.62%
Female:	(1299.7)	38.62	(1879.4)	45.43	(2002.7)	57.37
Chi-square:	19.02		38.01		10.99	
P-value:	.0002		<.0001		.0030	
Odds ratio:	0.62		0.58		0.62	
95% CI:	(0.50, 0.77)		(0.49, 0.69)		(0.48, 0.80)	

CMH chi-square: 42.55 (p<.0001)

*Weighted sum

表 6.6 下半部分列出了考虑调查设计后使用 SUDAAN 中的 CROSSTAB 程序进行相同分析的结果。在 PROC 命令句式中，DESIGN = wr 指定有放回抽样，表示并没有使用有限总体校正。NEST 命令指定阶层和 PSU 变量，WEIGHT 命令给出了权重变量。SUBGROUP 命令指出了 3 个离散变量，而 LEVELS 命令则明确了每个离散变量的(取值)类别的数量。TABLES 命令界定了列联表的形式。PRINT 命令要求报告 nsum(频数)，wsum(加权后频数)，

rowper(行百分比),cor(粗比率比),upcor(cor 的上限),lowcor(cor 的下限),chisq(卡方统计值),chisqp(卡方统计值的 p 值),cmh(CMH 统计值),cmhpval(CMH 的 p 值)。

维他命使用的加权后百分比与未加权百分比稍微不同。在表格上半部的三个不同的分析中,除了在中间那个教育水平之外,Wald 卡方值比 Pearson 卡方值小。虽然在更低的教育水平中比数比几乎保持不变,但在中间及更高的教育水平中却降低了。SUDAAN 中的 CROSSTAB 程序没有计算普通比数比,但可从 logistic 回归分析中得到,这将在下文中讨论。

Logistic 回归分析

前面讲到的线性回归分析对很多社会科学家来说可能用处不大,因为在社会科学研究中很多变量通常都用类别(名义的或者次序的)进行测量。有一系列用于分析分类数据的统计方法,从前一节所提到的基本的跨列表分析到具各种不同联结函数的广义线性模型等等。正如 Knoke 和 Burke(1980)所观察到的,建模的方法在社会科学研究中改革了列联表分析方法,废除了大部分用于研究离散类别测量的变量之间的关系的旧方法。有两种方法被社会科学家广泛应用:使用最大似然方法的对数线性模型(Knoke & Burke, 1980; Swafford, 1980)和加权后最小二乘法(Forthofer 7 Lehnen, 1981; Grizzled, Starmer & Koch, 1969)。这两种用于分析复杂调查数据的方法在本书的第一版中有所阐述。在这些模型中,单元比例或者它们的函数(如对数线性模型中的自然对数)用列联表中的(各种因素的)效应的线性组合表示。由于这些模型受限于列联表,因此在分析中无法包含连续的自变量。

在过去的十年间,社会科学家已经开始更加频繁地使用 logistic 回归分析,因为它可以把更多的解释性变量涵盖进来,其中就包括连续变量(Aldrich & Nelson, 1984; DeMaris, 1992; Hosmer & Lemeshow, 1989; Liao, 1994)。Logistic 回归和其他具不同联结函数的广义线性模型现在已经嵌入到复杂调查数据分析的(统计)软件包里了。调查分析者可以从一系列模型中选择最合适

的模型。用 Stata 和 SUDAAN，logit 模型的应用阐释如下。

方程 6.1 代表普通线性回归分析，检验的是一个连续因变量与一个或多个自变量之间的关系。Logistic 回归是一种用于检验一个类别因变量与一系列自变量之间关系的方法。下面这个方程表示对一个二分类结果变量 Y 建立模型的方式，$i = 1, 2, \cdots, n$：

$$\log[\pi_i/1-\pi_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{p-1} x_{p-1,\,i} \qquad [6.2]$$

在方程 6.2 中，π_i 是 $y_i = 1$ 的概率。这是一个具对数比或 logit 联结函数的广义线性模型。这里用最大似然的方法(Eliason，1993)来估计参数，而不是最小二乘估计法。因为这里要解的联立方程对参数并不是线性的，所以使用的是迭代技术。最大似然理论同时还可估计这些参数(β)的估计值的协方差矩阵，前提是个体观察值是随机且相互独立的。

正如方差模型的分析一样，如果一个变量有 l 个类别，那么我们在模型中只用 $l-1$ 个类别。我们将从被忽略类别(或者参照组)的效应中测量这 $l-1$ 个类别的效应。β 的估计值($\hat{\beta}$)是模型中对应类别与被忽略类别之间的 logit 的差别，即模型中对应类别与被忽略类别的比数比的自然对数。因此，$e^{\hat{\beta}}$ 就等于对模型中的其他变量进行调整后得到的比数比。Logistic 回归结果通常以比数比的方式总结和解释(Liao，1994，第 3 章)。

在复杂调查数据中，需要对最大似然估计进行调整，因为每个观察值都有一个抽样权重。结合了权重的最大似然解通常被称为 pseudo 或者加权后最大似然估计(Chambless & Boyle，1985；Roberts，Rao & Kumar，1987)。虽然点估计是用 pseudo 似然方法计算的，但 $\hat{\beta}$ 的协方差矩阵却是用第 4 章讨论的其中一种方法估计的。正如前面提到的，与这个协方差矩阵相关的自由度大约等于 PSU 的数量减去阶层的数量。因此，标准的模型拟合的似然比检验不能用于调查数据的 logistic 回归分析之中。相反要使用调整后的 Wald 检验。

在 logistic 回归模型中，对合适预测变量的选择或纳入与线性回归中的过程类似。当分析的是一个大型调查数据库时，前一节所描述的前期分析策略在为 logistic 回归分析做准备时非常有用。

为阐释 logistic 回归分析，我们用 Stata 对表 6.6 用到的数据进行分析。分析结果见表 6.7。我们稍微编辑了一下 Stata 输出结果以适应表格大小。结果变量是维他命使用，解释变量包括性别（1 = 男性；0 = 女性）和教育程度。这个模型基于表 6.6 所显示的 CMH 统计值，没有包含交互项。首先，我们进行了标

表 6.7　NHANES III 阶段 II 中美国成人性别和教育程度对维他命使用影响的 logistic 回归分析 (n = 9920)：用 Stata 进行的分析

(A) 标准 logistic 回归（忽略设计效应和权重）

```
. xi: logit vituse i.male i.edu

        i.male            _Imale_0-1        (naturally coded; _Imale_0 omitted)
        i.edu             _Iedu_1-3         (naturally coded; _Iedu_1 omitted)
        Iteration 0:   log likelihood = -6608.3602
        Iteration 1:   log likelihood = -6445.8981
        Iteration 2:   log likelihood = -6445.544
        Iteration 3:   log likelihood = -6445.544
Logit estimates                                    Number of obs   =      9920
                                                   LR chi2(3)      =    325.63
                                                   Prob > chi2     =    0.0000
Log likelihood = -6445.544                         Pseudo R2       =    0.0246
------------------------------------------------------------------------------------------
     vituse |   Coef.  Std. Err.      z     P>|z|   [95% Conf.Int.] Odds Ratio  [95% Conf.Int.]
------------+-----------------------------------------------------------------------------
   _Imale_1 | -.4418    .0427    -10.34    0.000   -.5256  -.3580     .6429     .5912  .6990
    _Iedu_2 |  .2580    .0503      5.12    0.000    .1593   .3566    1.2943    1.1773 1.4285
    _Iedu_3 |  .7459    .0512     14.56    0.000    .6455   .8462    2.1082    1.9069 2.3308
      _cons | -.5759    .0382    -15.07    0.000   -.6508  -.5010
------------------------------------------------------------------------------------------
```

(B) 拟合优度检验

```
. lfit

Logistic model for vit, goodness-of-fit test
       number of observations =      9920
 number of covariate patterns =         6
             Pearson chi2(2) =       0.16
                 Prob > chi2 =     0.9246
```

(C) 利用权重和设计因素的调查 logistic 回归

```
. svyset [pweight=wgt], strata (stra) psu(psu)
. xi: svylogit vituse i.male i.edu

        i.male            _Imale_0-1        (naturally coded; _Imale_0 omitted)
        i.edu             _Iedu_1-3         (naturally coded; _Iedu_1 omitted)

Survey logistic regression
pweight:  wgt                                       Number of obs    =      9920
Strata:   stra                                      Number of strata =        23
PSU:      psu                                       Number of PSUs   =        46
                                                    Population size  =   9920.06
                                                    F(   3,     21)  =     63.61
                                                    Prob > F         =    0.0000
------------------------------------------------------------------------------------------------
     vituse |   Coef.  Std. Err.     t    P>|t|   [95% Conf. Int.]     Deff   Odds Rat.[95% Conf. Int.]
------------+-----------------------------------------------------------------------------------
   _Imale_1 | -.4998    .0584    -8.56   0.000   -.6206   -.3791    1.9655    .6066    .5376   .6845
    _Iedu_2 |  .2497    .0864     2.89   0.008    .0710    .4283    2.4531   1.2836   1.0736  1.5347
    _Iedu_3 |  .7724    .0888     8.69   0.000    .5885    .9562    2.8431   2.1649   1.8013  2.6019
      _cons | -.4527    .0773    -5.86   0.000   -.6126   -.2929    2.8257
------------------------------------------------------------------------------------------------
```

(D) 检验系数的线性合并

```
. lincom _Imale_1+_Iedu_3, or

 ( 1)   _Imale_1 + _Iedu_3 = 0.0
-----------------------------------------------------------------------------
      vituse | Odds Ratio   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
         (1) |   1.3132      .1518       2.36    0.027      1.0340     1.6681
-----------------------------------------------------------------------------
```

准的 logistic 回归分析，忽略了权重和设计特征。结果列于表格 A 部分。在 logit 命令之前加上 xi 并在相应的变量名前面加上 i.，Stata 会自动对离散变量执行效果或二分类编码。

表中的输出结果列出了每个离散变量被忽略的类别。在这个例子中，"男性"这个类别，在模型中它的效应是从参照组"女性"的效应中测量出来的。至于教育程度，低于高中毕业是参照类别。似然比卡方值等于 325.63(自由度 $df = 3$)，p 值为 < 0.00001，我们否定性别与教育程度这两种因素一起对维他命使用没有影响这个假设，表示存在着显著影响。但 pseudo R^2 表示维他命使用的大部分变异情况不能被这两个变量所解释。性别和教育的参数估计值、它们的标准误差估计值以及相应的检验统计值也列于表中，所有的因素都显著。

在模型命令中加入 or 会得到比数比，而不是 β 系数。估计到的男性的比数为 0.64，表示在对教育进行了调整之后，男性在使用维他命方面的比数[①]是女性比数的 64%。这个比数比与表 6.6 中的 CMH 普通比数比相同。比数比的显著性可以用 z 检验或者置信区间进行检验。教育程度的第三个类别的比数比表示，具大学教育程度的人其使用维他命的可能性是那些具相同性别但教育年限少于 12 的人的两倍。没有任何一个置信区间包含 1，表示所有的影响都是显著的。

表 6.7 的 B 部分列出了拟合优度统计值(自由度 $df = 2$ 的卡方值)。p 值很大表示主效应模型适合数据情况(与饱和模型没有显著差别)。

在这种简单的情况下，与模型拟合优度相关的 2 个自由度也可以解释为与性别—教育交互项相关的 2 个自由度。因此，关于维他命的使用比例，并不存在性别与教育的交互作用，这也证实了表 1.15 中的 CMH 分析结果。

表 6.7 的 C 部分显示了把调查设计考虑在内时，对相同数据进行 logistic 回归分析的结果。对数似然值没有列出来，原因在于使用的是 pseudo 似然法。这里我们用的是 F 统计值而不是似然比统计值。同样，p 值表示主效应模型是对零模型

---

① 即：使用的概率/不使用的概率。——译者注

的一个显著的改进。由于抽样权重的原因，估计到的参数和比数比有变化。正如在设计效应中反映出来的一样，β 系数的标准误差增加了。虽然标准误差增加了，但性别和教育的 β 系数显著地不等于 0。对教育调整后的男性的比数从 0.64 下降到 0.61。虽然对第二个教育水平保持不变，但 p 值却显著变大了，从 <0.0001 上升到 0.008，原因在于我们把设计(因素)也考虑在内了。

在运行了 logistic 回归之后，D 部分显示了对参数线性组合的效应进行检验的结果。我们想要检验男性的参数与第三个教育水平的参数之和等于 0 这个假设。由于没有交互效应，因此得到的 1.3 这个比数比可以解释为，具大学教育的男性其使用维他命的比数相对于参照组(少于 12 年教育的女性)的比数高 30%。SUDAAN 也用于执行 logistic 回归分析，在单机版中使用 LOGISTIC 程序或者在 SAS 的可赎版本(不同的用以与 SAS 中的标准 logistic 程序区分开来的名字)中使用 RLOGIST 程序。

最后，logistic 回归模型还可对综合估计建立预测模型。由于大部分健康调查是为估计全国性的统计值而设计的，因此我们很难对小区域进行健康特征的估计。得到小地区估计值的一种方法就是利用全国性的健康调查和地区性的人口信息进行综合估计。LaVange，Lafata，Koch 和 Shah(1996)利用一个拟合了 NHIS(美国国家健康访问调查)和 ARF(地区资源档案)的 logistic 回归模型对美国各州各县老人活动受限的普遍情况进行估计。由于 NHIS 以复杂调查设计为基础，他们用 SUDAAN 对 NHIS 的活动受限预测因素以及 ARF 中的县级变量进行拟合，建立了 logistic 回归模型。然后，以模型为基础的概率估计值被用于计算小地区的活动受限估计值。

其他 Logistic 回归模型

上面所讨论的二分类 logistic 回归模型可被扩展用以处理多于 2 个回应类别的情况。这种回应类别中有些是次序的，如感知的健康状况：很好，好，一般和差。有些可能是名义上的，比如在宗教崇拜中。这些次序的和名义上的结果变量可被当做是一系列离散和连续自变量的函数来进行检验。利用 Stata 或者

SUDAAN，这种建模方法可用于复杂调查数据。在这一部分，我们举了两个这种分析的例子，但没有对它们进行深入讨论和分析。关于这种方法的详细描述及分析请见 Liao(1994)。

为阐释次序 logistic 回归模型，我们检验了基于 BMI 的肥胖类别。公共卫生营养学家们用下面的标准把 BMI 分成几种肥胖程度：极度肥胖（$BMI \geqslant 30$），超重（$25 \leqslant BMI < 30$），正常（$18.5 \leqslant BMI < 25$），偏瘦（$BMI < 18.5$）。基于 NHANES III 阶段 II 的数据，18%的美国成年人为极度肥胖，34%超重，45%正常，3%偏瘦。我们想要检验这 4 种肥胖程度（bmi2：1=极度肥胖，2=超重，3=正常，4=偏瘦）与一系列解释变量包括年龄（连续变量）、教育、黑人和西班牙裔等之间的关系。

对于这 4 个肥胖的有序类别，我们对以下 3 组概率建模使之成为解释变量的函数：

Pr{极度肥胖}相对于 Pr{其他所有类别}

Pr{极度肥胖+超重}相对于 Pr{正常+偏瘦}

Pr{极度肥胖+超重+正常}相对于 Pr{偏瘦}

然后 3 个二分类 logistic 回归模型可分别用于上述每一组比较。但是，鉴于这个肥胖类别的自然排序，我们可以在成比例发生比假设的基础上，同时考虑这 3 个二分类模型来估计解释变量的"平均"效应。这里假设的是，每个不同结果类别的回归线之间是相互平行的，但截距可以不同（这个假设需要用卡方统计值进行检验，检验结果没有列于表中）。下面这个方程代表了 $j = 1, 2, \cdots, c-1$（c 是因变量的类别数量）的模型：

$$\log\left(\frac{\Pr(category \leqslant j)}{\Pr[category \geqslant (j+1)]}\right) = \alpha_j + \sum_{i=1}^{p} \beta_i x_i \qquad [6.3]$$

从这个模型，我们估计 $(c-1)$ 个截距项和一系列 $\hat{\beta}$。

表 6.8 显示了以上用 SUDAAN 进行的分析结果。SUDAAN 命令列于表格顶部。第一个命令 PROC MULTILOG 明确了所用的程序。DESIGN，NEST 和 WEIGHT 的具体说明与表 6.6 的一样。REFLEVEL 说明教育的第一个类别作为参照组（如果不明确指出则变量的最后一个类别作为参照组）。

表 6.8 NHANES III 阶段 II 中美国成年人教育、年龄和种族对肥胖程度影响的次序 logistic 回归分析($n = 9920$)：用 SUDAAN 进行的分析

```
proc multilog design=wr;
nest stra psu;
weight wgt;
reflevel edu=1;
subgroup bmi2 edu;
levels    4     3;
model bmi2=age edu black hispanic/ cumlogit;
setenv decwidth=5;
run;

Independence parameters have converged in 4 iterations
-2*Normalized Log-Likelihood with Intercepts Only: 21125.58
-2*Normalized Log-Likelihood Full Model          : 20791.73
Approximate Chi-Square (-2*Log-L Ratio)          :   333.86
Degrees of Freedom                               :        5

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable: BMI2
----------------------------------------------------------------------
BMI2 (cum-logit),
  Independent                                                 P-value
  Variables and             Beta                              T-Test
  Effects                   Coeff.      SE Beta   T-Test B=0  B=0
----------------------------------------------------------------------
BMI2 (cum-logit)
  Intercept 1             -2.27467      0.11649    -19.52721  0.00000
  Intercept 2             -0.62169      0.10851     -5.72914  0.00001
  Intercept 3              2.85489      0.11598     24.61634  0.00000
AGE                        0.01500      0.00150      9.98780  0.00000
EDU
  1                        0.00000      0.00000        .        .
  2                        0.15904      0.10206      1.55836  0.13280
  3                       -0.20020      0.09437     -2.12143  0.04488
BLACK                      0.49696      0.08333      5.96393  0.00000
HISPANIC                   0.55709      0.06771      8.22744  0.00000
----------------------------------------------------------------------

       -------------------------------------------------------
       Contrast                Degrees of             P-value
                               Freedom      Wald F    Wald F
       -------------------------------------------------------
       OVERALL MODEL           8.00000    377.97992   0.00000
       MODEL MINUS
         INTERCEPT             5.00000     36.82064   0.00000
       AGE                     1.00000     99.75615   0.00000
       EDU                     2.00000     11.13045   0.00042
       BLACK                   1.00000     35.56845   0.00000
       HISPANIC                1.00000     67.69069   0.00000
       -------------------------------------------------------
-----------------------------------------------------------
BMI2 (cum-logit),
  Independent
  Variables and                          Lower 95%    Upper 95%
  Effects                 Odds Ratio     Limit OR     Limit OR
-----------------------------------------------------------
AGE                        1.01511       1.01196      1.01827
EDU
  1                        1.00000       1.00000      1.00000
  2                        1.17239       0.94925      1.44798
  3                        0.81857       0.67340      0.99503
BLACK                      1.64372       1.38346      1.95295
HISPANIC                   1.74559       1.51743      2.00805
-----------------------------------------------------------
```

分类变量列在 SUBGROUP 命令之后，这些变量每个所包含类别的数量列在 LEVELS 命令之后。MODEL 这一行命令列出了因变量，随后是一系列自变量。MODEL 命令中的关键词 CUMLOGIT 指明用的是成比例发生比模型。若没有这个关键词，SUDAAN 将会拟合多类别logistic 回归模型，这个模型会在下文进行讨论。最后，SETENV 命令要求在发表结果时报告 5 个小数点。

输出结果显示了三个截距项和自变量的一组系数估计结果。第二个表框中的结果表明所有主效应都显著。第三个表框中的比率比可以用与二分类 logistic 回归中相同的方式进行解释。在控制了其他自变量之后，西班牙裔比非西班牙裔具有 1.7 倍甚至更高的属于极度肥胖的比率。在解释这些结果之前，我们必须检查是否满足了成比例发生比假设，但是输出结果并没有给出任何检验这个假设的统计值。为检验这个假设，我们运行了三个普通 logistic 回归分析(极度肥胖相对所有其他，极度肥胖＋超重相对正常＋偏瘦，极度肥胖＋超重＋正常相对偏瘦)。年龄对应的三个比率比分别为 1.005，1.012 和 1.002，它们与表 6.8 底部所显示的 1.015 很接近。其他自变量对应的比率比也相当类似。我们认为成比例发生比假设看来可以接受。

利用 svyolog 程序，Stata 还可用于拟合成比例发生比模型，但它所拟合的模型稍微有所不同。在方程 6.3 中，一组 $\beta_i x_i$ 项加在截距项之后，但在 Stata 模型中它被去掉了。因此，从 Stata 中估计到的 β 系数其符号与从 SUDAAN 中得到的相反，而绝对值则相同。这意味着从 Stata 中得到的比数比是从 SUDAAN 中得到的比数比的倒数。这两种方法给出的截距项估计值是完全相同的。Stata 用的术语是 cut 而不是 intercept。

对名义的结果类别，可以使用多类别 logistic 回归模型。用这个模型，我们可以检验一个多类别名义结果变量(没有顺序的)与一组解释变量之间的关系。这个模型指定结果变量中的某一个类别为基准类别，并估计相对于这个基准变量，处于第 j 个类别的概率的比数的对数。这个比数称为相对风险，而这个比数的对数则称为广义 logit。

我们使用了上面用到的相同的肥胖程度。虽然我们明显地看到肥胖程度具有顺序，我们在此把它视为名义变量，因为我们想拿其他肥胖程度与正常这

个类别进行比较。相应地，我们把肥胖程度的编码做了修改(bmi3：1＝极度肥胖，2＝超重，3＝偏瘦，4＝正常[基准类别])。我们使用了 3 个预测变量，包括年龄(连续变量)、性别(1＝男性[参照组]，2＝女性)和目前的吸烟状态(1＝现在抽烟，2＝从未抽过[参照组]，3＝以前抽过现在不抽)。下面的方程代表了这个模型：

$$\log\left(\frac{\Pr(obese)}{\Pr(normal)}\right)=\beta_{0,1}+\beta_{1,1}(age)+\beta_{2,1}(male)+\beta_{3,1}(c.smoker)+\beta_{4,1}(p.smoker)$$

$$\log\left(\frac{\Pr(overweight)}{\Pr(normal)}\right)=\beta_{0,2}+\beta_{1,2}(age)+\beta_{2,2}(male)+\beta_{3,2}(c.smoker)+\beta_{4,2}(p.smoker)$$

$$\log\left(\frac{\Pr(underweight)}{\Pr(normal)}\right)=\beta_{0,3}+\beta_{1,3}(age)+\beta_{2,3}(male)+\beta_{3,3}(c.smoker)+\beta_{4,3}(p.smoker) \qquad [6.4]$$

我们使用 SUDAAN 来拟合上面的模型，结果列于表 6.9 中(我们稍微编辑了一下 Stata 输出结果以适应表格大小)。SUDAAN 中所有的命令与前面成比例发生比模型中的命令类似，除了在 MODEL 这一行命令中省略了 CUMLOGIT。Stata 中的 svymlogit 程序也可用于拟合多类别回归模型。

表 6.9 同时显示了 β 系数和相对风险比率(标记为比数比)，同时也列出了标准误差和检验 $\beta=0$ 的 p 值。年龄在极度肥胖相对正常和超重相对正常的比较中都是一个显著的因素，但在偏瘦相对正常的比较中则不是。虽然性别在极度肥胖相对正常的比较中没有影响，但在另两个比较中却有。我们来看比数比的表格，相对于正常这个类别男性处于超重的相对风险比率是女性的两倍以上，给定年龄和抽烟状态相同的情况下。相对于正常这个类别，在保持年龄和性别不变的情况下，现在抽烟的人其处于极度肥胖的相对风险比率只有那些从不抽烟的人的 68%①。

① 原文中为 0.68%，实为有误。——译者注

表 6.9 NHANES Ⅲ 阶段 Ⅱ 中美国成人性别和抽烟状态对肥胖程度影响的多类别 logistic 回归分析 ($n = 9920$)：用 SUDAAN 进行的分析

```
proc multilog design=wr;
nest stra psu;
weight wgt;
reflevel csmok=2 sex=2;
subgroup bmi2 csmok sex;
levels   4      3      2;
model bmi3=age sex csmok;
setenv decwidth=5;
run;
```

```
Independence parameters have converged in 6 iterations
Approximate ChiSquare (-2*Log-L Ration)          :   587.42
Degrees of Freedom                               :       12

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Generalized Logit
Response variable: BMI3
```

BMI3 log-odds)		Independent Variables and Effects				
		Intercept	AGE	SEX = 1	CSMOK = 1	CSMOK = 3
1 vs 4	Beta Coeff.	-1.33334	0.01380	0.08788	-0.39015	-0.27372
	SE Beta	0.14439	0.00214	0.12509	0.07203	0.13206
	T-Test B=0	-9.23436	6.43935	0.70251	-5.41617	2.07277
	P-value	0.00000	0.00000	0.48941	0.00002	0.04958
2 vs 4	Beta Coeff.	-1.25883	0.01527	0.76668	-0.24271	-0.02006
	SE Beta	0.13437	0.00200	0.08275	0.11067	0.09403
	T-Test B=0	-9.36835	7.64830	9.26512	-2.19307	-0.21335
	P-value	0.00000	0.00000	0.00000	0.03868	0.83293
3 vs 4	Beta Coeff.	-2.07305	-0.01090	-1.16777	0.33434	0.04694
	SE Beta	0.48136	0.00742	0.25280	0.30495	0.26168
	T-Test B=0	-4.30663	-1.46824	-4.61937	1.09637	0.17936
	P-value	0.00026	0.15558	0.00012	0.28427	0.85923

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	15.00000	191.94379	0.00000
MODEL MINUS INTERCEP	12.00000	68.70758	0.00000
INTERCEPT	.	.	.
AGE	3.00000	22.97518	0.00000
SEX	3.00000	64.83438	0.00000
CSMOK	6.00000	6.08630	0.00063

BMI3(log-odds)		Independent Variables and Effects				
		Intercept	AGE	SEX = 1	CSMOK = 1	CSMOK = 3
1 vs 4	Odds Ratio	0.26360	1.01390	1.09186	0.67695	0.76054
	Lower 95% Limit	0.19553	1.00941	0.84291	0.58323	0.57874
	Upper 95% Limit	0.35535	1.01840	1.41433	0.78573	0.99946
2 vs 4	Odds Ratio	0.28399	1.01539	2.15260	0.78450	0.98014
	Lower 95% Limit	0.21507	1.01120	1.81393	0.62397	0.80689
	Upper 95% Limit	0.37499	1.01959	2.55450	0.98633	1.19059
3 vs 4	Odds Ratio	0.12580	0.98916	0.31106	1.39702	1.04805
	Lower 95% Limit	0.04648	0.97408	0.18439	0.74341	0.60994
	Upper 95% Limit	0.34052	1.00447	0.52476	2.62526	1.80086

现有的软件也支持其他可用于分析复杂调查数据的统计模型。比如，SUDAAN 支持用于生存分析的 Cox 的回归模型（比例风险模型），虽然截面调

查很少提供纵向数据。利用 SUDAAN、Stata 和其他软件支持的程序，其他由不同的联结函数定义的广义线性模型也可应用到复杂调查数据中。

基于设计和基于模型的分析

到目前为止，所有的分析都依赖于基于设计的方法，因为分析中结合了抽样权重和设计特征。在把这些分析与基于模型的方法联系起来之前，我们先简单看看用于这些分析的调查数据。NHANES III 一共产生 2812 个 PSU，覆盖整个美国。这些 PSU 有时包括单个县，有时包括 2 个或以上相邻的县。这些县是行政单位，而调查并不是为了对这些单位分别进行估计而设计的。从这些单位中，81 个 PSU 被选中，选择概率与 PSU 的规模成比例——13 个来自确定的阶层。此外，从根据人口统计学特征而非地理位置形成的 34 个阶层中每层抽出 2 个 PSU。同样地，阶层并不是以定义人口参数为目的而设计的。抽样的第二个阶段涉及地区图块，包括城市或者城郊街区或者其他邻近的地理区域。具有更多的少数种族人口的图块用更高的概率进行选取。抽样的第三个阶段涉及把抽中的地区图块里面的所有家庭罗列出来，然后按照以图块特征为基础的比率对它们进行抽样。抽样的第四个阶段是对抽中的家庭里面的个人进行抽样以进行访问。这些次级单位用以辅助抽样而不是用于界定人口参数。公开的数据档案只包括阶层和 PSU，支持次级抽样单位的标识并不包括在内。抽样权重的计算基于被访问个人的选择概率，此外还包括对无应答和事后分层的加权调整。前几章所讨论的许多分析问题是由大型社会和健康调查的执行方式造成的。现有的数据无法支持分层线性模型以融合多阶段选择设计。

由于受访个体的不等选择概率加上无应答和事后分层调整，若在 NHANES III 数据的描述性分析中忽略抽样权重，就会难以令人信服。在描述性分析中进行加权的原因已经很清楚。正如表 6.2 显示的，在未加权估计中年龄和与种族相关的变量的偏差很大。对所有的变量，加权后估计值的标准误差与未加权时的结果很类似，这意味着抽样权重的变异情况并不会增加方差。但是，当把 PSU 和阶层考虑进来时，加权后估计值的标准误差相当大，正如设计效应所反

映的。降低方差的一种方法是使用包含辅助信息的模型。典型的例子是熟悉比率和回归估计(Cochran, 1977,第 6 章)。在这些估计中,辅助信息采取的形式是已知的与目标变量相关的伴随变量的人口均值。辅助信息的使用可以扩展为分布函数的估计(Rao, Kovar & Mantel, 1990),但辅助信息在例行描述性分析中的使用是有限的,因为很难找到合适的辅助信息。

回归分析中模型的使用是显而易见的。表 6.4 中的未加权估计严格地以基于模型的方法为基础。它忽略了抽样权重和设计特征,但许多相关的设计信息已经包含在自变量之中:比如年龄(对年老者的过抽样)以及黑人和西班牙裔(对少数人口的过抽样)。其实,基于模型的系数估计值与加权后估计值类似,意味着基于模型的分析在这个情况下很合理。一个显著的例外是教育的系数,它在两种估计方法间非常不同。在基于模型的分析中教育非常不显著,但在加权后分析中却非常显著。教育的效应无法被基于模型的分析检测到,因为对年长者而言,教育的作用在逐渐消失。虽然年龄已包含在模型中,但年龄和教育的交互作用却没有包含在模型中。这个例子说明,抽样权重的使用避免了模型的错误设定。

Korn 和 Graubard(1995b)利用 1988 年美国全国母婴健康调查数据,更进一步地阐释了在回归分析中使用抽样权重的好处。这个调查对低出生体重的婴儿进行了过抽样。估计到的胎龄对出生体重影响的回归线在未加权分析和加权后分析间非常不同。虽然未加权拟合均等反映了样本观察值且不对总体进行描述,但加权拟合把回归线拉到了估计到的总体所处的位置。这两个变量之间的关系其实是曲线的。相反,如果我们使用二次回归,那么未加权和加权后回归就会更具一致性。

正如上面对表 6.4 的分析结果所进行的讨论,对加权后和未加权回归间的差别进行细致的检验有时可以找到应该加入到方程中的重要变量或者交互项。未加权和加权后估计之间的差异表明,把抽样设计考虑在内可以避免可能出现的总体模型的错误设定。已有文献已经提出了几种用于检验加权后和未加权估计之间差异的统计值(DuMouchel & Duncan, 1983; Fuller, 1984; Nordberg, 1989)。Korn 和 Graubard(1995a)利用基于设计的变异性把这些检验统

计值应用到 NHANES I 和 II 的数据中。他们建议当低效性很小时，使用基于设计的分析。否则，附加的建模假设可以融入到分析中去。他们已经注意到次级抽样单位并不对公众开放这个问题，也已指出需要在大型健康调查中增加 PSU 的数量。这些检验受限于点估计，因此他们的结论不能应用到所有的情况中。Pfeffermann(1993, 1996)提供了更详细的关于这些问题和与之相关问题的讨论。

基于设计的分析避免可能的模型错误设定这个事实，说明利用 SUDAAN, Stata 和其他复杂调查分析软件的分析适合 NHANES 数据。甚至在基于设计的分析中，我们也使用了模型来具体说明所感兴趣的参数，但推论把抽样权重考虑在内了。这种情况下，基于设计的分析可称为模型辅助方法(Sarndal, Swensson & Wretman, 1992)。基于设计的理论依赖于大的抽样规模来对参数做推论。对小样本来说基于模型的分析可能是更好的选择。在数据收集中没有使用概率抽样时，不存在应用基于设计的推论的基础。当实质理论和先前经验调查支持所提议的模型时，基于模型的方法可能更有道理。

相比在回归分析中，基于模型的分析的构思在列联表分析中更不易理解。我们已经讨论了把抽样方案考虑在内时基于设计的分析的基本原理。正如在回归分析中一样，很有必要关注加权后比例与未加权比例之间的差别。如果存在显著差别，我们就应该检查为什么会存在差别。在表 6.6 中，未加权比例和加权后比例很接近，但对高中毕业生和具大学教育的人来说，维他命使用和性别的加权后比数比比未加权比数比稍微小一些，而对教育程度低于高中毕业的人来说，加权后和未加权比数比则差不多相同。两个较高教育水平的轻微差别可能来自种族或其他因素。如果未加权和加权后比数比的差别大很多且归因于种族，我们就应该对不同的种族群体分别检验这种相关性。在列联表中考虑附加因素的作用可以通过使用 logistic 回归模型实现。

Logistic 回归中模型的使用以及相关问题与线性回归中的一模一样。对加权后和未加权分析进行细致检验可以提供有用的信息。在表 6.7 中，加权后和未加权的系数估计值类似。我们发现加权对截距项的影响要大于对系数的影响。表 6.7 所示的分析是用 logistic 回归分析数据的一个简单的演示，并没有特

别考虑对合适模型的选择。我们并没有对表 6.8 中次序 logistic 回归模型和表 6.9 中多类别 logistic 回归模型进行类似的、没有使用权重和抽样设计的、基于模型的分析，因为我们并没有被设定出一个合适的包含所有相关自变量的模型。

综上所述，复杂调查数据同时需要基于模型和基于设计的分析。基于设计的方法产生近乎无偏的估计值或者相关性，但标准误差可能是低效的。基于模型的方法需要在选择模型时作出假设，而错误的假设会导致相关性和标准误差的有偏估计。

第 7 章 | 总结

在本书中，我们讨论了调查数据分析存在问题的各方面以及用于处理复杂抽样设计的使用所带来的问题的方法。重点关注的是对各种问题的理解和处理办法的逻辑，而不是技术操作指导。我们还介绍了准备复杂调查数据分析的实用指引，阐释了某些可用于执行各种分析的软件的使用。用于复杂调查分析的软件现在很容易找到，而且随着个人计算机计算能力的增强，许多复杂的分析方法也能很容易地执行。然而，数据分析者需要明确设计方法、对某些分析创建复合权重和对调查分析选择合适的检验统计值。因此，使用者应该对抽样设计以及相关的分析问题有很好的理解。

虽然对这些议题所介绍的材料主要是针对调查数据分析者，但我们希望这部分内容能同时激励调查设计者和数据制造者更加关注调查数据使用者的需要。调查数据最初的收集仅仅只是为了记数，随着这些数据越来越多地用于分析性研究，调查设计者必须考虑把某些与设计相关的信息包括在数据之中，以实现更加合适的分析，同时减轻数据使用者的负担。除了提供阶层和抽样单位的代码，数据制造者还应该创建与设计相应的抽样权重，甚至还要包括结合了无应答和事后分层调整的复合权重。

最后，我们必须指出，我们一直是站在基于设计的统计推论立场之上的，同时也简要介绍了另一种方法，即基于模型的推论。这两种方法各有优势。简单来说，基于模型的推论假定样本是从概念上的超总体中得到的一组便利的观察值。设定好的模型下的总体参数是主要关注点，而抽样选择方案则没有推论那么重要。因此，在这里抽样设计的角色变得不重要了，统计估计使用的是这个设定好的模型下的预测方法。自然地，如果模型设定是错误的，那么这种估计

就会产生偏差，而且即使在大样本中偏差也可能是巨大的。基于设计的推论要求把抽样设计考虑在内，这是传统的方法。有限总体是主要关注点，分析目标在于在重复的抽样中找到设计无偏的估计值。

我们相信，当推论基于抽样数据时抽样设计是非常重要的，相比那些更具可预测性的物理现象，社会现象中的描述尤其重要。同时，我们需要对模型的恰当性进行评估，分析性模型的作用必须在所有数据分析中得到重视。同时使用设计和模型的推论相比只单独使用其中一种的推论可能更加成功。我们还相信，这两种不同的方法是互补的，联合使用两者会有益处。Brewer（1995，1999）和 Sundberg（1994）也陈述过类似的观点。[7]我们既不能忽略那些在社会调查设计中普遍存在的实际性问题，也不能忽略那些一直持续的无应答及其他引起非抽样误差的问题。这些实际性问题和社会科学研究中实质理论的现状迫使我们现在只能更多地依赖传统的方法。基于模型的方法应该在消除调查设计和分析间的鸿沟这个问题上提供额外的思考方法。

已有大量的理论和实际文献探讨了复杂调查数据的分析问题。除了本书所列的参考文献，另有关于本书所讨论问题及相关议题的、更深入的、独立成册的分析（Korn & Graubard，1999；Lehtonen & Pahkinen，1995；Skinner，Holt & Smith，1989）。在对复杂调查数据进行二手分析时，分析者应该意识到这种分析很可能会遭到抽样设计局限甚至抽样设计错误所导致的严重破坏。掌握有限抽样信息的分析者很有必要向有经验的专业调查技术员进行咨询。

注释

[1] 比率估计的方法用于估计两个变量的总体比率(比如,水果重量与产生果汁多少的比率)。它还用于获得某个变量的更加准确的估计值(比如现在的收入 y),通过构建其与另一个密切相关的变量的比率(比如,上次普查时的收入 x)。这个样本比率(y/x,或收入的变化)接着被应用到上次普查时的收入来获得现在收入的估计值,它比没有使用辅助变量进行的估计更加准确。详细介绍见 Cochran(1977,第 6 章)。

[2] Holt 和 Smith(1979)把事后分层描述为一种稳健估计技术。依据条件分布,他们发现自加权后的样本均值通常是有偏差的,而事后分层则针对极端的样本构造为更准确的估计提供了保护。他们认为在现有的状况中,事后分层在抽样调查中没有得到充分应得的考虑。此外,事后分层还可以针对由抽样选择中的无应答和其他问题带来的任何异常现场为更准确的估计提供保护。

[3] SUPER CARP 和 PC CARP 可从爱荷华州立大学统计实验室获得,而且对偏统计的用户很有用。美国人口普查局的 Fay 博士使 CPLX 变得可用,它对用调整后的 BRR 以及折叠法进行离散多变量分析很有帮助(Fay, 1985)。CENVAR 和 VPLX 程序现在可从美国人口普查局找到。由 CDC(美国疾病控制和预防中心)开发的用于流行病学和统计学分析的 Epi Info 系统包括了用于复杂调查数据分析的 CSAMPLE 程序。密歇根大学社会研究所的 OSIRIS 统计软件系统包含了一些用于描述性统计和回归分析的程序(只对大型电脑开放)。还有其他用于特殊调查项目如世界生育率调查(CLUSTERS)的程序。另外有两个用于调查数据分析的 SAS 宏系列程序,包括 GES(可从加拿大统计局获得)和 CLAN(可从瑞典统计局获得)。

[4] 美国国民健康与营养调查(NHANES Ⅲ)是由美国国家卫生统计中心(NCHS)组织的一系列持续的调查,用以评估美国人口的健康与营养状况。NHANES 已经完成了好几轮调查。NHANES Ⅰ是在 1971—1973 年开展的,NHANES Ⅱ完成于 1976—1980 年,而 NHANES Ⅲ则在 1988—1994 年进行。1982—1984 年间他们还组织了一次专门针对西班牙裔人口的调查(西班牙 NHANES)。NHANES 已经变成一个持续的调查项目,数据现在每 2 年公开一次(1999—2000, 2001—2002, 2003—2004)。NHANES 从数目庞大的被访者中,通过个体访问和包括诊断检测与其他临床实践中用到的程序的健康检查,收集各种与健康相关的信息(S. S. Smith, 1996)。NHANES 的设计比较复杂,用以适应经费和调查要求的实际限制,最终得到了一个包含家庭中符合条件个人的分层多阶段概率集群样本(NCHS, 1994)。其初级抽样单位(PSU)是县或相邻县的小组群,随后的分层抽样单位包括普查小区、家庭集群、家庭和符合条件的个人。对学前儿童、老人或者穷人采取了过取样,以提供足够的这些子群体的人数。公开的微观数据库所包含的抽样权重是扩展权重(根据无应答和事后分层而调整过的选择概率的倒数)。NHANES Ⅲ分两个阶段执行。多阶段抽样设计得到了 89 个抽样地区,这些地区接着又随机分成两组:其中 44 个点在 1988—1991 年间调查,剩下的 45 个在 1991—1994 年间调查。每个阶段的样本都可视为独立样本,而联合样本则可用于大型分析。

[5] 热卡填补(就近填补)的方法借用数据库中其他观察值的取值。有许多不同的选择捐赠观察值的方法。通常是按照选中的人口特征变量及其他变量如阶层和 PSU 等对数据进行排序来建立填补单元。然后从与具缺失值的观察值处于同一个单元的其他观察之中选取捐赠者。通过输入个体取值而非均值,这种方法在很大程度上避免对方差的低估。这种方法广泛用于美国人口普查局及其他调查机构。详情请见 Levy 和 Lemeshow(1999:409—411)。

[6] 自由度为 1 的 Wald 统计值基本上等于均值为 0 的正态变量的平方除以其标准差。对于涉及自

由度大于 1 的假设，Wald 统计值是正态变量的平方的矩阵扩展。

[7] Brewer 和 Mellor(1973)精辟地阐述了基于设计和基于模型的方法之间的差别。Brewer(1995)充分地描述了不同方法的联合使用，这种联合使用方法进一步被应用到分层相对分层对称抽样的例子中(Brewer, 1999)。Sundberg(1994)结合方差估计表达了类似观点。Graubard 和 Korn(2002)回顾和阐释了从基于设计的抽样设计转到基于模型的抽样设计的优点以及抽样推论的优点。他们指出，在分层抽样中，把 FPC 因素从标准的基于设计的方差公式中去掉，以获得合适的方差公式作基于模型的推论是不充分的。在集群抽样中，标准的基于设计的方差公式会严重低估基于模型的变异情况，就算在最后(抽样)单位的抽样比例很小的情况下也如是。他们得出结论认为，基于设计的推论是一个有效的、合理不受模型干扰的方法，可对有限总体参数作出推论，但他们建议对基于设计的方差估计进行简单的修正，以对超总体参数做一些模型假设情况下的推论，而这也常常引起主要的科学关注。

参考文献

Aldrich. J. H. , & Nelson, F. D. (1984). *Linear probability, logit, and probit models* (Quantitative Applications in the Social Sciences, 07—045). Beverly Hilis, CA: Sage.

Alexander. C. H. (1987). A model-based justification for survey weights. *Proceedings of the Section of Survey Research Methods* (American Statistical Association), 183—188.

Bean, J. A. (1975). *Distribution and properties oj variance estimation for complex multistage probability samples* (Vital and Health Statistics, Series 2[65]). Washington, DC: National Center for Health Statistics.

Binder. D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. *51*, 279—292.

Brewer, K. R. W. (1995). Combining design-based and model-based infereoce. In B. G. Cox. D. A. Binder, B. N. Chinnappa. A. Christianson, M. J. Colledge, & P. S. KOll(Ed,.), *Business Survey methods*, 589—606. New York: John Wiley.

Brewer, K. R. W. (1999). Design-based or prediction-based inference? Stratified random vs. stratified balanced sampling. *International Statistical Review*, *67*, 35—47.

Brewer, K. R. W. , & Mellor, R. W. (1973). The effect of sample structure on analytical surveys. *Australian Journal of Statistics*. *15*. 145—152.

Brick, J. M. , & Kalton. G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*, 215—238.

Bryk. A. S. , & Raudenbush. S. W. (J992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park. CA: Sage.

Chambless, L. E. , & Boyle, K. E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics-Theory and Methods*, *14*, 1377—1392.

Chao, M. T. , & Lo, S. H. (1985). A bootstrap method for finite populalions. *Sankhya*. *47*(A). 399—405.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley.

Cohen, S. B. (1997). An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *The American Statistician*. *51*, 285—292.

Davis, J. A. , & Smith, T. W. (1985). *General Social Survey*, 1972—1985: *Cumulative codebook* (NORC edition). Chicago: National Opinion Research Center, University of Chicago and the Roper Center. University of Connecticut.

DeMaris, A. (1992). *Logit modeling* (Quantitative Applications in the Social Sciences. 07—086), Thousand Oaks, CA: Sage.

Deming. W. E. (1960). *Sample design in business research*. New York: John Wiley.

DuMouchel, W. H. , & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, *78*, 535—543.

Durbin. J. (1959). A note on the application of Quenouille's method of bias reduction to the estima-

tion of ratios. *Biometrika*. *46*, 477—480.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1—26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eliason. S. R. (1993). *Maximum likelihood estimation: Logic and practice* (Quantitative Applications in the Social Sciences, 07—096). Beverly Hills, CA: Sage.

Eltinge, J. L., Parsons, V. L., & Jang, D. S. (1997). Differences between complex-design-based and II D-based analyses of survey data: Examples from Phase I of NHANES III. *Stats*, *19*, 3—9.

Fay, R. E. (1985). A jackknife chi-square test for complex samples. *Journal of the American Statistical Association*, *80*, 148—157.

Flyer, P., & Mohadjer, L(1988), *The WesVar procedure*. Rockville, MD: Westat.

Forthofer. R. N., & Lehnen. R. G. (1981). *Public program analysis: A categoricai data approach*. Belmont, CA: Lifetime Learning Publications.

Frankel, M. R. (1971). *Inference from survey samples*. Ann Arbor: Institute of Social Research, University of Michigan.

Fuller, W. A, (1975). Regression analysis for sample surveys, *Sankhya*, *37*(C), 117—132.

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, *10*, 97—118.

Goldstein. H., & Silver, R. (1989). Multilevel and multivariate models in survey analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith(Eds.), *Analysis of complex survey data* (pp. 221—235). New York: John Wiley.

Goodman, L. A. (1972). A general model for the analysis of surveys. *American Journal of Sociology*, *77*, 1035—1086.

Graubard, B. I., & Korn, E. L(1996), Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, *5*, 263—281.

Graubard, B. I., & Korn, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science*, *17*, 73—96.

Grizzle, J. E., Starmer, C F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*. *25*. 489—504.

Gurney, M., & Jewett. R. S. (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association*. *70*, 819—821.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Sratistical Association*. *78*, 776—807.

Heitjan, D. F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, *87*(4), 548—550.

Hinkins, S., Oh, H. L., & Scheuren, F. (1994). Inverse sampling design algorithms. *Proceed-*

ings of the Section on Survey Research Methods (American Statistical Association), 626—631.

Holt. D., & Smith, T. M. F. (1979). Poststratification. *Journal of the Royal Statistical Society*, *142* (A), 33—46.

Holt, D., Smith. T. M. F., & Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, *143* (A). 474—487.

Horton, N. J., & Lipsitz. S. R. (2001). Multiple imputation in practice: Comparison of software packages for regession models with missing variables. *The American Statistician*. *55* (3). 244—254.

Hosmer, D. W., & Lemeshow. S. (1989). *Applied logistic regression*. New York: John Wiley.

Judkins, D. R. (1990). Fay method of variance estimation. *Official Statistics*, *6* (3), 233—239.

Kalton, G. (1983). *Introduction to survey sampling* (Quantitative Applications in the Social Sciences. 07—035). Beverly Hills. CA: Sage.

Kalton, G., & Kasprszky, D, (1986), The treatment of missing survey data. *Survey Methodology*, *12* (1), 1—16.

Kendall, P. A.. & Lazarsfeld, P. F. (1950). Problems of survey analysis. In R. K. Merton & P. F. Lazarsfeld (Eds.), *Continuities in social research: Studies in the scope and method of "The American soldier."* New York: Free Press.

Kiecolt, K. J., & Nathan, L. E. (1985). *Secondary analysis of survey data* (Quantitarive Applications in the Social Sciences, 07—053). Beverly Hills, CA: Sage.

Kish. L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, *44*, 380—387.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Kish, L., & Frankel, M. R. (1974). Inferences from complex samples. *Joumal of the Royal Statistical Society*, *36* (B), 1—37.

Knoke, D., & Burke, P. J. (1980). *Log-linear models* (Quantitative Applications in the Social Sciences, 07—020). Beverly Hills, CA: Sage.

Koch, G. G., Freeman, D. H., & Freeman, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *international Statistical Review*, *43*, 59—78.

Konijn. H. (1962). Regression analysis in sample surveys. *Journal of the American Statistical Association*, *57*, 590—605.

Korn, E. L., & Graubard, B. I. (1995a). Analysis of large health surveys: Accounting for the sample design. *Journal of the Royal Statistical Society*, *158* (A). 263—295.

Korn, E. L., & Graubard, B. I. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, *49*, 291—295.

Korn. E. L., & Graubard. B. I. (1998). Scatterplots with survey data. *The American Statistician*, *52*, 58—69.

Korn. E. L., & Graubard, B. I. (1999). *Analysis of health surveys*. New York: John Wiley.

Korn. E. L., & Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society*, *8* (65, pt. 1), 175—190.

Kott, P. S. (1991). A model-based look at linear regression with survey data. *The American Statis-*

rician, *45*, 107—112.

Kovar. J. G., Rao, J. N. K., & Wu, C. F. J. (1988), Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, *16* (Suppl.), 25—45.

Krewski, D., & Rao, J. N. K. (1981), Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, *9*, 1010—1019.

LaVange, L. M., Lafata. J. E.,. Koch, G, G., & Shah. B. V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research*, *5*, 311—329.

Lee, E. S., Forthofer. R, N., Holzer, C. E., & Taube. C. A. (1986). Complex survey data analysis: Estimation of standard errors using pseudo-strata. *Journal of Economic and Social Measurement*, *14*, 135—144.

Lee, E. S., Forthofer. R. N., & Lorimor, R. J. (1986). Analysis of complex sample survey data: Problems and strategies. *Sociological Methods and Research*, *15*, 69—100.

Lee, K. H. (1972). The use of partially balanced designs for the half-sample replication method of variance estimation. *Journal of the American Statistical Association*, *67*. 324—334.

Lehtonen. R., & Pahkinen. E. J. (1995). *Practical methods for design and analysis of complex surveys*. New York: John Wiley.

Lemeshow. S., & Levy. P. S. (1979). Estimating the variance of ratio estimates in complex surveys with two primary sampling units per stratum. *Journal of Statistical Computing and Simulation*, *13*. 191—205.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*, New York: John Wiley.

Levy. P. S., & Stolte, K. (2000). Statistical methods in public health and epidemiology: A look at the recent past and projections for the future. *Statistical Methods in Medical Research*, *9*, 41—55.

Liao. T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models* (Quantitative Applications in the Social Sciences. 07—101). Beverly Hills, CA, Sage.

Little. R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Little. R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.

Lohr. S. L. (1999). *Sampling: Design and analysis*. New York: Duxbury.

McCarthy. P. J. (1966). *Replication: An approach to the analysis of data from complex surveys* (Vital and Health Statistics. Series 2[14]). Washington, DC: National Center for Health Statistics.

Murthy. M. N., & Sethi, V. K. (1965). Self-weighting design at tabulation stage. *Sankhya*. *27* (B). 201—210.

Nathan, G., & Holt, D. (1980). The effects of survey design on regression analysis. *Journal of the Royal Statistical Society*. *42* (B), 377—386.

National Center for Health Statistics(NCHS). (1994). *Plan and operation of the Third National*

Health and Nutrition Examination Survey, 1988—1994(Vital and Health Statistics, Series 1 [32]). Washington, DC: Government Printing Office.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review*. *71*, 593—627.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*. *5*, 223—239.

Pfeffermann. D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*. *61*, 317—337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, *5*, 239—261.

Pfeffermann, D., & Homes, D. J. (1985). Robustness considerations in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*. *148* (A). 268—278.

Pfeffermann. D., & Nathan, G. (1981). Regression analysis of data from a cluster sample. *Journal of the American Statistical Association*, *76*, 681—689.

Plackett R. L., & Burman, P. J. (1946). 'The design of optimum multi-factorial experiments. *Biometrika*. *33*, 305—325.

Quenouille, M. H. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society*. *11* (B), 68—84.

Rao, J. N. K., Kovar, J. G., & Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary infonnation. *Biometrika*, *77*, 365—375.

Rao, J. N. K., & Scott, A. J. (1984). On chi-square tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*. *12*, 46—60.

Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex: survey data. *Journal of the American Statistical Association*. *83*, 231—241.

Rao, J. N. K., Wu. C. F. J., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, *18*(3). 209—217.

Roberts. G., Rao. J. N. K., & Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, *74*, 1—12.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*. 377—387.

Royall. R. M. (1973). *The prediction approach to finite population sampling theory: Application to the hospital discharge survey*(Vital and Health Statistics, Series 2[55]). Washington, DC: National Center for Health Statistics.

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*. *5*, 283—310.

Sarndal. C. E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*. *5*, 25—52.

Sarndal. C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Shah, B. V., Holt, M. H., & Folsom, R. E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, *47*, 43—57.

Sitter, R. R. (1992). Resampling procedure for complex survey data. *journal of the American Statistical Association*, *87*. 755—765.

Skinner, C. J., Holt, D., & Smith. T. M. F. (Eds.). (1989). *Analysis of complex survey data*. New York: John Wiley.

Smith, S. S. (1996). The third National Health and Nutrition Examination Survey: Measuring and monitoring the health of the nation. *Stats*, *16*, 9—11.

Smith, T. M. F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, *139*(A), 183—204.

Smith. T. M. F. (1983). On the validity of inferences on non-random samples. *Journal of the Royal Statistical Society*. *146*(A), 394—403.

Sribney, W. M. (1998). Two-way contingency tables for survey or clustered data. *Stata Technical Bulletin*. *45*, 33—49.

Stanek. E. J., & Lemeshow, S. (1977). The behavior of balanced half-sample variance estimates for linear and combined ratio estimates when strata are paired to form pseudo strata. *American Statistical Association Proceedings: Social Statistics Section*, 837—842.

Stephan, F. F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, *43*, 12—39.

Sudman. S. (1976). *Applied sampling*. New York: Academic Press.

Sugden. R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in sampling inference. *Biometrika*, *71*. 495—506.

Sundberg. R. (1994). Precision estimation in sample survey inference: A criterion for choice between various estimators. *Biometrika*, *81*, 157—172.

Swafford, M. (1980). Three parametric techniques for contingency table analysis: Non-technical commentary. *American Sociological Review*. *45*, 604—690.

Tepping, B. J. (1968). Variance estimation in complex surveys. *American Statistical Association Proceedings*, *Social Statistics Section*. 11—18.

Tukey. J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Tukey, J. W. (1986). Sunset salvo. *The American Statistician*, 40, 72—76.

U. S. Bureau of the Census. (1986, April). *Estimates of the population of the United States*, *by age*, *sex*, *and race*. 1980 to 1985 (Current Population Reports. Series P—25, No. 985). Washington, DC: Author.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Woodruff. R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statstical Association*, *66*, 411—414.

Zhang. P. (2003). Multiple imputation: Theory and method. *International Statistical Review*. *71*, 581—592.

译名对照表	
Approximation	近似值
attrition-adjusted	损耗调整
attrition rate	损耗率
auxiliary variable	辅助变量
balanced repeated replication	对称重复抽样
body mass index (BMI)	体重指数
bootstrap method	自主抽样法
circular systematic sampling scheme	循环系统抽样方案
cluster sampling	集群抽样
collapsed strata	折叠层
confidence interval	置信区间
contingency table	列联表
covariance	协方差
Cox proportional hazards model	Cox 比例风险模型
cross-sectional surveys	截面调查
cross-tabulation analysis	跨列表分析
design effect	设计效应
design-weighted least squares (DWLS)	设计加权最小二乘法
disproportionate stratified sample design	不成比例分层抽样设计
effect coding	效果编码
expansion estimation	扩展估计
expasion weight	扩展权重
extreme value	极端值
finite population correction factor	有限总体校正因素
first-order derivative	一阶导数
follow-up survey	追踪调查
generalized linear models	广义线性模型
general social survey(GSS)	综合社会调查
goodness of fit	拟合优度
host sample	主样本
hot deck imputation	热卡填补(就近填补)
independent and identically distributed	独立同分布
indicator variable	指示变量
intraclass correlation coefficient, ICC	组内相关系数
inverse sampling design algorithm	逆抽样设计演算法

item nonresponse	选项无应答
jackknife repeated replication	"折叠式"重复抽样
joint inclusion probabilities	联合被选中概率
linear combination	线性组合
linearization method	线性化方法
log odds	对数比
logistic regression	logistic 回归
log-likelihood	对数似然
log-linear regression	对数线性回归
longitudinal data	纵向数据
main effects model	主效应模型
mean imputation	均值填补
mean square error	均方差
measurement error	测量误差
median-trace plot	中位数标示图
missing value	缺失值
multicollinearity	多重共线性
multinomial logistic regression	多类别 logistic 回归
multiple imputation	多重填补
multiple-layered nesting	多层嵌套
multistage cluster sampling	多阶段集群抽样
national opinion research center(NORC)	全国民意研究中心
nonprobability sampling	非概率抽样
null model	零模型
odds ratio	比数比
optimum(Neyman) allocation	最优(内曼)分配
ordered logistic regression	次序 logistic 回归
ordinary least squares(OLS) estimation	普通最小二乘估计
orthogonal matrix	正交矩阵
oversampling	过取样
paired selection design	配对选择设计
parameter	参数
partially balanced replicate	部分对称复合样本
point estimate	点估计
poisson regression	泊松回归
population mean	总体均值

population total	全及总体
poststratification adjustment	事后分层调整
PPS(probability proportional to size) sampling	PPS 抽样(按规模大小成比例概率抽样)
poverty index	贫困指数
presicion	精确度
primary sampling unit(PSU)	初级抽样单位
probability distribution	概率分布
probability sampling	概率抽样
propagation of variance	方差传递法
proportional hazard model	比例风险模型
proportional odds assumption	成比例发生比假设
proportional odds model	成比例发生比模型
pseudo-replication	拟复合抽样
pseudo-strata	拟阶层
quota sampling	配额抽样
random variable	随机变量
ratio estimate	比率估计
regression imputation	回归填补
relative weight	相对权重
repeated systematic sampling	重复系统抽样
replicate weights	复合权重
replicated sampling	复合抽样
response rate	应答率
robust estimation	稳健估计
sample design	抽样设计
sample mean	样本均值
sample weights	抽样权重
sampling fraction	抽样比例
sampling frame	抽样框
saturated model	饱和模型
scatterplot	散点图
selection bias	选择偏差
self-weighting	自加权
side-by-side boxplot	对比箱形图
simple random sampling	简单随机抽样
simple random sampling with replacement SRSWR	简单有放回随机抽样

simple random sampling without replacement SRSWOR	简单无放回随机抽样
simple two-stage cluster sampling	简单二阶段集群抽样
simultaneous equation	联立方程
standard error	标准误差
statistical inference	统计推断
stratified multistage cluster sampling	分层多阶段集群抽样
stratified random sampling	分层随机抽样
stratified sampling	分层抽样
survival analysis	生存分析
synthetic estimation	综合估计
systematic sampling	系统抽样
systolic blood pressure(SBP)	收缩压
target population	目标总体
Taylor series method	泰勒级数法
the best unbiased predictor	最优无偏估计
ultimate cluster approximation	最终集群逼近法
unequal probability design	不等概率设计
unequal-sized cluster	不同规模集群
unit nonresponse	单元无应答
variance-covariance matrix	方差—协方差矩阵
variance estimation	方差估计
weighted average	加权平均数
weighted fitting	加权拟合
weighted least-square	加权最小二乘法
weighted least-square modeling	加权最小二乘建模
weighted odds ratios	加权后对数比
weighted proportions	加权后比例
weighted sum	加权总和

缺失数据

作 者 简 介

保罗·D.阿利森(Paul D. Allison)

美国宾夕法尼亚大学社会学教授。曾于1976年在威斯康星大学取得博士学位,之后在芝加哥大学及宾夕法尼亚大学做统计学的博士后研究。关于社会科学中的统计方法,他已出版5本专著,发表文章超过25篇。近期出版的书籍有《多元回归:一个最重要的方法》(*Multiple Regression*: *A Premier*)和《使用SAS系统的logistic回归:理论与应用》(*Logistic Regression Using the SAS System*: *Theory and Application*)。主要研究兴趣为科学家的职业与利他行为理论。

译 者 简 介

林毓玲

曾先后获得台湾大学学士学位和美国芝加哥大学硕士学位,现为香港科技大学硕士研究生。研究兴趣为社会分层和流动。

序

在经验社会科学研究中缺失数据的问题很普遍，大多非实验性研究所报告的统计结果都立足于较小的样本数，有时比初始选择个案数目还要小。在一些变量上相对多缺失观察值会降低有效的 N。假设有个意见调查，多变量分析中有效样本只有原来的一半，这种情况在现实中并不少见。假设商学院的 Mary Rose 教授在一个消费者态度及行为调查中检验一个 $N = 1000$ 受访者的概率样本。她使用一般计算机选项成列删除（也就是任一受访者有缺失任一模型变量即被排除），对支出估计一个合理设定的多元回归模型。结果实际可得的个案降至 $N = 499$。这就产生了严重的问题。这 499 个受访者是否仍"代表"了总体？要拒绝零假设，样本是否太小？为了保持样本数，是否应该尝试成对删除？抑或，有其他新的方法值得考虑？这些问题及其他问题都在 Paul Allison 这本杰出的专题著作中讨论到。

"观察值是随机缺失的"，这是根据留下的个案以面对处理数据缺失时的通常论点。但这个假设是隐含的。假若观察值"完全随机缺失"，这表示没有任何变量，不论是因变量（Y）或自变量（X），其缺失分数都不与该变量自身的值相关。例如，上述的支出变量，对于支出多者其未回答率应不比支出少者的未回答率高。对于其他模型变量，假设在相同的条件下，则 499 个次样本将可代表一次科学的抽取，允许有效的推论。而且，它允许回归估计值是不偏及一致的。无问题派的研究者可能喜欢这种 MCAR 随机性，但这需要有很强的假设来支持。

较为实际一点的假设为观察值是"随机缺失的"（MAR）。假设在控制了其他变量后，如果 Y 的值不能预设缺失分数的位置，则 Y 变量缺失数据为随机的。所以在上述的举例中，职业地位（X）可能与支出的缺失数据相关，高地位的受访者更可能低报支出。一旦 X 在右手边，那么 Y 的观察值将会是随机的。在 MAR 情况下，如 Allison 所言，缺失数据产生机制是可忽略的。虽然他也论及不可忽略的缺失数据机制的困难细节，但他这本专题著作着重于在 MAR 条件下，以改良估计处理的方法。

如果数据是 MAR，则估计的质量很大程度地取决于系统性误差的位置。令人鼓舞的是，当相关缺失数据仅限于自变量时，则成列删除仍能产生不偏的估计值。例如，在例

子中，职业地位 X 缺失数据可能与另一个自变量年龄(Z)相关；例如：没有报告年龄的可能年纪较大且地位较高。在年龄较大与报告支出没有相关的条件下，则没有误差。事实上，正如 Allison 巧妙论证的，在一些 MAR 情况下，标准成列删除选项比传统缺失数据修正方法(成对删除，虚拟变量调整或平均值替换)表现更好。

处理缺失数据问题的新策略占用了本专题论著的大部分篇幅。在缺失数据的条件下回顾最大似然估计，即 ML 估计，他以一个仔细筛选的美国大专院校毕业率的数据为例，解释了插补法的 EM 算法。后几章超越了 ML 方法，解释多重插补方法，并讨论了不可忽略的缺失数据。这本书是最新的处理缺失数据的精心杰作，几乎所有的统计书籍都很少涉及这个主题。Paul Allison 也睿智地提醒我们，缺失数据最佳的解决方法是“没有任何最佳解决方法”。但如果你也有这个问题且在寻求补救方法，那么就请阅读本书的内容。

——迈可 S. 路易贝克(Michael S. Lewis-Beck)

第 1 章 | 导论

任何做统计分析的人早晚都会遇到缺失数据的问题。在一个典型的数据组中，对于某些个案信息是缺失的。例如，在要求个人报告其收入的调查中，通常很大比例的受访者会拒绝回答。彻底拒绝只是造成缺失数据的一个原因。而在自己填答的调查中，人们通常会漏看或忘记回答一些问题。即使是专业的调查员偶尔也会忽略某些问题。有时候受访者说他们只是不知道如何回答或者是没有可用的信息，而有时候某些问题对于一些受访者是不适用的，如让未婚者评价他们的婚姻质量。在长期的研究中，这一波被访问的人在下一波调查前可能会死亡或搬走。当数据从多个行政记录中收集得来时，有些记录可能也会不慎缺失。

因为有这些和其他许多的原因，导致缺失数据在社会和健康科学两者中，成为了一个普遍存在的问题。为什么它会是个问题呢？因为几乎所有标准统计方法都假设每个个案具有可用于分析中的所有变量的信息。确实，大多数的统计教科书没有提到任何有关缺失数据或如何处理缺失数据的信息。

一个众所周知、通常为统计软件默认的、简单的解决方式是：在分析中当某个案的任何变量具缺失数据时，便简单地将该个案从分析中排除。结果便得到一个没有缺失数据的数据组，可以使用任何传统方法分析。这个策略通常被社会科学称之为成列删除或个案删除，但有时也会被称为完整个案分析。

除了简单以外，成列删除也有一些吸引人的统计特质，这将会在后面讨论。对任何使用过"成列删除"的人而言，它也有一个明显的重大缺点：在很多运用中，成列删除会排除原始样本的一个很大的比例。例如，假使你已收集了一个样本为 1000 人的数据且想要估计一个有 20 个变量的多元回归。每一个变量都

有 5%的个案具缺失数据，且每一个变量数据缺失的机会与任何其他变量信息缺失的机会是独立的。那么可预计只有 360 个个案具完整数据，丢弃了其他 640 个“个案”。如果你仅从某网站上下载数据，你可能就不会感觉太糟糕，虽然你也希望能有更多的个案。但如果你已对 1000 人中的每个人都花了 200 美元进行访问，就可能会非常懊悔，因为大概有 130000 美元浪费了(至少对于这个分析是如此)。但在实际操作中，确实有方法可以从这 640 个不完整的个案中抢救回某些东西，因为在这 640 个不完整的个案中，许多个案可能只缺少 20 个变量中的其中一个信息。

许多备选方法被提出来，且其中数个“方法”将在本书中被重新讨论。但遗憾的是，这些方法大多没有价值，且很多未必优于成列删除。虽然这些方法背后的理论已经至少有十年之久，但却仅在过去数年间才可以在计算机上操作。甚至到目前为止，多重插补及最大似然估计仍需要花费大量的时间与精力用于学习方法并根据例行程序执行它们。但如果你想要把事情做对，你通常要付出代价。

最大似然和成列删除两者都有我们希望所能达到的好的统计特性。然而，必须注意的是，这些方法和其他所有的方法一样，其效度基于某些容易被违反的假设。不仅如此，而且没有方法可以检验大部分重要的假设是否被满足。虽然某些解决缺失数据的方法明显优于其他方法，但却没有一个可以被认为确实是好的，也不存在唯一真正解决缺失数据的好的方法。所以在设计和执行研究计划时，必须尽力使缺失信息的发生最小化。因为统计调整没有办法补救草率的研究。

第 2 章 | 假设

研究者通常会试着使那些在某一特别变量上有缺失值的个案与其他有观察值的个案变得没有差异。例如他们通常会提供证据说明报告和不报告其收入的人们在其他各变量上没有显著差异。更普遍地，研究者通常声称或假设他们的数据是随机缺失的，却没有完全理解这代表什么意思。统计学者过去甚至也曾对此概念感到模糊或模棱两可。然而，Rubin (1976)通过严格定义那些关于缺失数据机制不同的合理假设，将这些事物立足于一个坚固的基础上。虽然他的定义相当具技术性，但在此我将试着给出一个通俗的理解。

完全随机缺失的

假设一个特殊变量 Y 有缺失数据。如果 Y 数据缺失的概率与 Y 本身的值或在该数据组中任何其他变量的值都无关的话，那么 Y 的数据可以说是完全随机缺失的(MCAR)。当这个假设满足所有变量，有完整数据的个人组可被视为原本观察值组中的一个简单随机次样本。要注意 MCAR 不考虑 Y 的“缺失”与其他某个变量 X 的“缺失”相关之概率。例如，即使拒绝报告其年龄的人们总是拒绝报告他们的收入，但该数据仍然可能是完全随机缺失的。

如果平均而言不报告其收入的人们比那些报告收入者年轻，那么 MCAR 假设会被违反。很容易就可检验这个暗示，通过将该样本分为两组——报告收入者和未报告收入者，再检验他们的平均年龄的差异。如果实际上有数据呈现者和有数据缺失者两者间在所有观察变量上都没有系统性的差异，则该数据可以算是随机观察的。另一方面，仅由数据通过这个检验并不能说明 MCAR 假

设被满足了。还必须保证某一特殊变量上的缺失与该变量的值没有关系。

虽然 MCAR 是一个相当强的假设，但有时候它也是合理的，特别是当数据缺失是研究设计的一部分时。当某个特殊变量测量起来太过昂贵时，这样的设计通常很吸引人。相应的策略便是只针对较大样本中的某个随机次集合，测量这个昂贵的变量，这也意味着对剩余的样本而言，该数据是完全随机缺失的。

随机缺失的

一个合理性较弱的假设为数据是随机缺失的(MAR)。如果在分析中控制了其他变量后，Y 缺失数据的概率与 Y 值无关，则称 Y 的数据为随机缺失的。为更正式地表达，假设只有两个变量 Y 和 X，X 总是被观察到但 Y 有时会缺失。那么 MAR 指

$$\Pr(Y \text{ missing} \mid Y,\ X) = \Pr(Y \text{ missing} \mid X)$$

用文字表示，这个表达式意味着，在同时给定 Y 和 X 时，Y 缺失数据的条件式概率，等于在只单独给定 X 的条件下 Y 缺失数据的概率。例如，如果收入缺失数据的概率取决于婚姻状态，但在每一个婚姻状态类别中，缺失收入的概率与收入无关。一般而言，在控制了其他观察变量后，如果有数据缺失的那些个人相对于那些有数据呈现者，那么对于该变量倾向于较低(或较高)的值，数据不会是随机缺失的。

检验 MAR 条件是否被满足是不可能的，但在直觉上理由应该是很充分的。因为我们不知道缺失数据的值，所以我们无法比较有缺失值者和没有缺失值者，进而看他们是否在该变量上有系统性的差异。

可忽略的

如果(a)数据为 MAR，且(b)管制缺失数据过程的参数与要估计的参数无关，则缺失数据的机制是可忽略的。可忽略性基本上指不需要将缺失数据机制

模型化为估计过程中的一部分。然而，确实需要使用特别技术以有效地利用数据，因为在实际运用中难以想象条件(b)不被满足的情况，因此作者在本书中将MAR和可忽略性视为相等的条件。甚至在极少数条件(b)不被满足的情况下，假设可忽略性的方法仍然运作的一样好，但你可以通过将缺失数据机制模型化从而做得更好。

不可忽略的

如果数据不是MAR，我们则说缺失数据机制是不可忽略的。在这个例子中，通常缺失数据机制必须被模型化从而得到所关注参数的好的估计值。关于不可忽略缺失数据的一个广泛使用的方法为Heckman (1976)因变量有选择偏差的两阶段估计回归模型。遗憾的是，对于不可忽略缺失数据的有效估计，需要非常好的关于缺失数据过程本质的基础知识，因为数据没有包含信息告知什么模型是适当的，而且结果会对模型的选择尤其敏感。因为这些原因且因为不可忽略缺失数据的模型要求对每个运用必须相当专业化，因此这本书主要将重点放在可忽略的缺失数据上。在第7章，作者简短地分析了一些处理不可忽略的缺失数据的方法。在第3章，我们将会看到成列删除有一些非常吸引人的特性，关于某些类型的不可忽略缺失数据这些特性也会非常明显。

第3章 | 传统的方法

虽然有许多不同的方法被提议用来处理缺失数据，但其中只有一些获得了普遍欢迎。但这些被广泛使用的方法中没有任何一个明显优于成列删除。在这部分，我从最简单的方法开始，简短地回顾一些方法。在评估这些方法时，我将特别关注他们在回归分析中的表现（包含 logistic 回归及 Cox 回归），许多评论也适用于其他类型的分析。

成列删除

如前所述，成列删除通过从样本中删除所关注模型中的在任何变量上有缺失值的观察值，并通过运用传统分析完整数据组的方法来实现。成列删除有两个明显的优点：(1)它可以用于任何类型的统计分析，包括从结构方程模型到对数线性分析；(2)并不需要特别的运算方法。根据缺失数据机制，成列删除有一些吸引人的统计特性。明确地说，如果数据为 MCAR，则减少的样本将会是原样本的一个随机次样本。这意味着，对于所关注的任何参数，如果估计值对于完整的数据组（没有缺失数据）的估计值是无偏误的，那么对于成列删除的数据组也会是无偏误的。此外，由成列删除的数据组所获得的标准误及检验统计量也如同它们在完整数据组中的一样适当。

当然，因为所利用的信息较少，标准误在成列删除的数据组中通常会比较大。它们也会倾向大于（在本书后面叙述的）由最适当的方法所获得的标准误，但至少你不用担心因为缺失数据而导致推论错误——这是大多数常用方法的一个大问题。

另一方面，如果数据不是MCAR，而只是MAR，那么成列删除可能会产生有偏误的估计值。例如，如果教育缺失数据的概率取决于职业地位，那么对职业地位进行关于教育的回归将会产生一个有偏误的回归系数估计值。因此，一般而言，成列删除对于违反MCAR假设的情况并不是稳健的。但出乎意料的是，成列删除对在回归分析中的自变量间违反MAR时是最稳健的。更明确地讲，如果任何因变量缺失数据的概率不取决于自变量的值，则使用成列删除的回归估计值将会是无偏误的（如果所有一般回归模型假设都被满足的话）。[1]

例如，假设我们想估计一个回归模型以预测年储蓄，其中一个自变量为收入，有40%的数据是缺失的。进一步地假设收入缺失数据的概率取决于收入和教育年数两者，教育年数为模型中另一个自变量。只要缺失收入的概率不取决于储蓄，回归估计值将会是无偏误的（Little, 1992）。

为什么会这样呢？有一个重要的原因。对回归模型的自变量做非比例分层化抽样并不会使系数估计值产生偏误。一个仅与自变量值相关的缺失数据机制在本质上与分层化抽样相同，也就是说，个案以基于其他变量值的概率而被选择进入样本中。这个结论不只适用于线性回归模型，也适用logistic回归、Cox回归、泊松回归及其他。

事实上，对logistic回归而言，甚至在更广泛的条件下，成列删除都能给予有效的推论。如果任何变量缺失数据的概率取决于因变量的值，而不取决于任何其他自变量的值，则使用成列删除的logistic回归会产生一致的斜率系数估计值及其标准误（Vach, 1994）。然而，截距估计值将会是有偏误的。只有在当任何缺失数据的概率同时取决于因变量和自变量两者时，使用成列删除的logistic回归才会是有问题的。[2]

总而言之，成列删除并不是一个差的处理缺失数据的方法。虽然它没有利用所有可得的信息，但至少当数据是MCAR时，它给予了有效的推论。正如我们将看到的，这几乎已经比所有其他普遍处理缺失数据的方法好多了。最大似然和多元插补方法（于后面几节讨论）在许多情况下可能会比成列删除更好，但对于回归分析，违反MAR假设时，成列删除甚至比这些复杂的方法更加稳健。更明确的是，当某个特别的自变量缺失数据的概率取决于该变量（而非因变量）

时，成列删除可能会比最大似然和多元插补更好。

对于这些关于回归分析的成列删除的主张有一个重要的提示。即对于样本中的所有个案回归系数都被假设是相同的。如果回归系数在横跨总体的次集合时发生变化，则该样本任何非随机的限制（例如，经过成列删除）都会导致回归系数向其中一个或另一个次集合倾斜。当然，如果在回归参数中察觉到这样的变化，就应该对不同的次样本做不同的回归，或将适当的交互作用包含在模型中（Winship & Radbill，1994）。

成对删除

也被称为可得个案分析，成对删除作为一个简单的备选方法可用于许多线性模型，包括线性回归、因子分析及更复杂的结构方程模型。它是广为人知的，例如，一个线性回归可以通过其样本平均数和协方差矩阵，或相同的，通过平均数、标准差及相关矩阵进行估计。成对删除的原理是要通过所有可得的个案来计算这些描述统计的每一个。例如，计算两个变量 X 和 Z 之间的协方差，所有同时具备 X 和 Z 两者数据的个案都会被使用。一旦总和测量值被计算出来了，它们就可用于计算我们所关注的参数，如回归系数。

如何执行这个原则却显得模棱两可。当计算需要每一个变量其平均数的协方差时，只使用这两变量都有数据的个案来计算平均数吗？还是使用所有变量的数据都可得的个案呢？因为所有的变异都产生具相同特性的估计量，所以毋需考虑诸如此类的问题。一般的结论是，如果数据为 MCAR，成对删除就产生一致的参数估计值（且因此在大样本中接近无偏误）。另一方面，如果数据是 MAR，但不是随机被观察到的，估计值就可能会严重偏误。

如果数据确实是 MCAR，成对删除可能会比成列删除更有效，因为更多信息被利用了。所谓更有效，指的是成对删除有比成列删除更少的抽样变异（较小的真实标准误）。然而，这并不总是正确的。线性回归模型的分析和模拟研究都指出，当变量间的相关性普遍较低时，成对删除会产生更有效的估计值，然而，当变量间的相关性较高时，成列删除则更好（Glasser，1964；Haitovsky，

1968; Kim & Curry, 1977)。

成对删除的一个大问题是:由传统软件所产生的标准误和检验统计量估计是偏误的。这个问题的症状是,当你输入一协方差矩阵于回归程序中时,你还必须指明样本数以计算标准误。有些成对删除的程序使用有最多缺失数据的变量的个案数目,而有些程序则用个案的最小数目以计算每一个协方差。然而,没有一个数目是令人满意的。原则上,有可能可以得到标准误的一致估计值,但公式很复杂,且目前在任何商业统计软件都无法执行。[3]

成对删除第二个偶尔会发生的问题是,在小样本中,建构的协方差或相关矩阵可能不是"正定的",这也暗示着回归运算根本无法实行。由于存在这些困难以及其对 MCAR 的偏离相对敏感,因此成对删除通常不被建议为成列删除的备选方法。

虚拟变量调整

这里还有另一个针对回归分析中缺失预测值的非常简单且直觉上很吸引人的方法(Cohen & Cohen, 1985)。假设某变量 X 有一些缺失数据,X 为回归分析中数个自变量的其中一个。我们建立一个虚拟变量 D,如果 X 数据缺失等于 1,如果没有缺失等于 0。我们也建立一个变量 X^*,使得

$$X^* = \begin{cases} X, \text{当数据没有缺失时} \\ c, \text{当数据缺失时} \end{cases}$$

其中 c 可以是任何常数。我们回归因变量 Y 于 X^*、D 及其他在预设模型中的所有变量。这个技术,被称为虚拟变量调整或缺失指标方法,该方法可以很容易被延伸至超过一个自变量具缺失值的数据中。

虚拟变量调整方法明显的好处在于它使用了所有可用的关于缺失数据的信息。将用 c 值代替缺失数据视为插补并不恰当,因为 X^* 的系数不会因为 c 值的不同选择而改变。而且,此模型唯一一个取决于 c 值的选择的面向为缺失值指标 D 的系数。为便于解释,一个简单可选的 c 值是非缺失个案的 X 的平均

数。这样 D 的系数可以被解释为，在控制了模型中的其他变量的情况下，X 具缺失数据的个体其 Y 的预测值减去具 X 平均数的个体其 Y 的预测值。X^* 的系数可被视为在有 X 数据的次群体中 X 的效应的估计值。

遗憾的是，如 Jones(1996)所证明的，这个方法通常会产生有偏误的系数估计值。[4]一个简单的模拟就能说明这个问题。从具 3 个变量的正态分布中抽样产生 10000 个个案，这 3 个变量为 X、Y 和 Z。回归 Y 于 X 和 Z，对于每个自变量，其真实的系数为 1.0。不出意外，利用整个具万个个案的样本而得到的最小二乘回归系数(如表 3.1 第 1 栏所示)相当接近真实值。

表 3.1　使用三种模型之仿真数据回归

系　数	完整数据	成列删除	虚拟变量调整
X	0.98	0.96	1.28
Z	1.01	1.03	0.87
D			0.02

其次，再随机地使 Z 值有 1/2 的概率缺失。因为缺失数据的概率与其他任何变量不相关，所以该数据为 MCAR。表 3.1 第 2 栏显示成列删除产生的估计值相当接近那些没有数据缺失的结果。另一方面，虚拟变量调整方法的系数则明显有偏误——X 系数高而 Z 系数太低。

另外有一个密切相关的方法被提议用于回归分析中的类别自变量。诸如此类的变量可通过建立一组虚拟变量来有代表性地处理这个问题，除了参照组外，每一个类别都有一个变量。这个方法是简单地建立一个额外的类别及一个额外的虚拟变量，以表示在该类别变量中具缺失数据的个体。然而，我们还有一个直觉上吸引人但即使当数据为 MCAR 时仍有偏误的方法(Jones, 1996; Vach & Blettneer, 1991)。

插补

许多处理缺失数据的方法都归在插补方法的大标题之下，其基本原理是要

以某些合理的猜测插补或替代缺失值，然后再接着按没有缺失数据的情况进行分析。当然，有许多不同的插补缺失值的方法。最简单的可能是边际平均数插补：对给定某个变量的每一个缺失值，都用有数据个体的该变量的平均数代替。众所周知，这个方法会产生有偏误的方差及协方差的估计值(Haitovsky，1968)，因此通常应该避免使用。

一个比较好的方法是利用多元回归的方法使用其他变量的信息，这个方法通常被称为条件式平均数插补。假设我们要估计一个有着数个自变量的多元回归模型。其中一个自变量 X，部分个案有缺失数据。对于那些有完整数据的个案，我们回归 X 于所有其他自变量上。使用相应的估计方程，我们会得到预测值用于具 X 缺失数据的个案。这些值用来代替缺失数据，并且可以接着按没有缺失数据的情况进行分析。

当超过一个自变量有缺失数据而且一般主题有许多变异时，这个方法就会变得比较复杂。一般而言，如果插补是全然根据其他自变量(而非因变量)且数据为 MCAR 时，最小二乘系数是一致的，暗示在大样本中几乎无偏误(Gourieroux & Monfort，1981)。然而，它们并不是完全有效的。可通过加权最小二乘(Beale & Little，1975)或广义最小二乘(Gourieroux & Monfort，1981)获得改良的估计量。

遗憾的是，所有这些插补方法都面临一个根本的问题：即按照完整数据的情况分析插补数据会低估标准误、高估检验统计量。传统分析分法没有单纯地对这个实际情况(即插补过程涉及缺失值的不确定性)进行调整。[5]后面几节会介绍一个解决这些问题的插补方法。

总结

所有从具缺失数据的个案中挽救信息的一般方法都会明显地使事情变得更糟：它们会引进重大的偏误，使分析对 MCAR 的偏离更加敏感，并产生不正确的标准误(通常太低)。由于存在这些缺点，成列删除因而看起来不算太糟。但还有更好的可用的方法。在下一章中，将会介绍可用于许多一般模型化目标

的最大似然方法。在第 5 章和第 6 章，也会介绍几乎可用于任何设定的多重插补方法。如果数据是 MAR，那么这两种方法都有不一般的特性。原则上，这些方法可用于不可忽略的缺失数据中，但需要一个有关数据缺失过程的正确的模型——这通常难以得到。

第 4 章 | 最大似然

最大似然是统计估计非常普遍的方法,广泛用于处理许多困难的估计问题。大多数读者可能熟悉 ML 作为估计 logistic 回归模型的偏好方法。当误差项假设为正态分布时,普通最小二乘线性回归也是一个 ML 方法。结果 ML 特别有利于处理缺失数据问题。在这一章,我们首先回顾 ML 估计值的一些普遍特性。然后,将介绍缺失数据机制为可忽略的假设下,ML 估计的基本原则。这些原则会用一个简单的列联表来说明。此章其余的部分会给出更复杂的例子,目的是要根据多变量正态分布估计一个线性模型。

回顾最大似然估计法

ML 估计的基本原则是,选取那些若取值是真实的,就可最大化观察到事实上已被观察到的概率的值,来作为估计值。为了达到这个目的,我们首先需要一个以数据和未知参数两者的函数来表达数据的概率的公式。当观察值为独立(一般的假设)时,该样本的总体似然(概率)就是所有个别观察值的似然的乘积。

假设我们要估计一个参数 θ。如果 $f(y \mid \theta)$ 为给定某个 θ 值时,观察到 Y 单一值的概率,则一个有 n 个观察值的样本的似然值为

$$L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

其中 $\prod$ 是重复乘法运算的符号。当然,我们仍需要确切指明 $f(y \mid \theta)$ 是什么。例如,假设 Y 是一个二分类变量,编码为 1 或 0,且 θ 为 $Y = 1$ 的概率。则

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

一旦我们有了称为似然函数的 $L(\theta)$，就有许多方法确定尽可能使似然值最大化的 θ 值。

ML 估计量有许多令人满意的特性，已知在相当广泛的条件下，它们是一致的、渐近有效且渐近正态的（Agresti & Finlay，1997）。一致性意味着估计值在大样本中接近无偏误。有效性意味着真实的标准误至少和任何其他相应的估计量的标准误一样小。

渐近部分意指这项陈述只是接近真实的，且样本量越大则越接近。最后，渐近正态性意指在重复抽样中，估计值接近正态分布（同样，接近程度会随样本数增加而增加）。这说明使用一个正态表来建构置信区间或计算 p 值是恰当的。

有缺失数据的 ML

当某些观察值具缺失数据时怎么办？当缺失数据机制是可忽略的时（因此为 MAR），我们可以简单地通过加总所有缺失数据可能值的一般似然来获得似然。例如，假设我们要对一个具 n 个独立观察值的样本收集两个变量 X 和 Y 的数据。对于前 m 个观察值，我们观察到 X 和 Y 两者，但对于剩下的 $n-m$ 个观察值，我们仅能测量到 Y。对于有完整数据的单一观察值，我们用 $f(x, y \mid \theta)$ 表示其似然值，其中 θ 为一组支配 X 和 Y 分布的未知参数。假设 X 是离散的，一个具 X 缺失数据的个案其似然就是 Y 的“边际”分布：

$$g(y \mid \theta) = \sum_{x} f(x, y \mid \theta)$$

当 X 为连续的时，加总以积分代替。整个样本的似然便为

$$L(\theta) = \prod_{i=1}^{m} f(x_i, y_i \mid \theta) \prod_{i=m+1}^{n} g(y_i \mid \theta)$$

问题就变成了寻找尽可能使这个似然值最大化的 θ 值。许多方法可用于解决这

个最适化问题，后面我将介绍其中的一些方法。

当缺失数据的模式为单调时，ML 特别简单。在一个单调的形态中，可以用一个顺序安排变量，以使样本观察值的排列情况为：如果数据对于某特别变量有缺失的话，那么必然地对于排在这个变量之后的其他变量数据也是缺失的。

举一个包含有 4 个变量 X_1、X_2、X_3 和 X_4 的例子。X_1 没有缺失数据，$X_2$100%的个案有缺失数据。缺失了 X_2 的个案在 X_3 和 X_4 这 2 个变量中数据也有缺失。额外地有 20%的个案 X_3 和 X_4 有缺失数据，但 X_2 没有缺失数据。一个单调的形态通常出现于面板研究，其中个人在不同时点退出且不再回到研究/数据中。

如果只有一个变量有缺失数据，该形态必然是单调的。考虑一个只有 X 具缺失数据的包含 2 个变量的例子。$f(x, y)$的联合分布可被写成 $h(x \mid y)g(y)$，其中 $g(y)$为 Y 的边际分布(前面已定义过)且 $h(x \mid y)$ 为给定 Y 时 X 的条件式分布。这样我们可把似然值重写为

$$L(\lambda, \phi) = \prod_{i=1}^{m} h(x_i \mid y_i; \lambda) \prod_{i=1}^{n} g(y_i \mid \phi)$$

这个表达式与前一个表达式在两个重要方面有所不同。首先，第二个乘积是针对所有观察值的，而不只是那些有 X 缺失数据的观察值的。其次，参数已被分成两部分：λ 表示给定 Y 时 X 的条件式分布，而 ϕ 则表示 Y 的边际分布。这些改变意味着我们可以分别最大化似然值的这两个部分。因此，如果 X 和 Y 有一个二变量正态分布，我们就可以计算整个样本的 Y 的平均数和方差。再者，对于有 X 数据的个案，我们可以回归于 Y 上。得到的参数估计值可以结合起来以产生我们可能会关注的任何其他参数的 ML 估计值，如回归系数。

列联表数据

这些 ML 估计的特征可以用列联表数据非常具体地说明。假设一个简单随机样本 200 人，我们要测量 2 个取值可能为 1 或 2 的二分类变量 X 和 Y。有 150 个个案，我们观察到 X 和 Y，并将获得的数据列于下面的列联表中：

	$Y=1$	$Y=2$
$X=1$	52	21
$X=2$	34	43

对于另 50 个个案，X 为缺失的，我们只观察到 Y：确切地说我们有 19 个个案 $Y=1$，31 个个案 $Y=2$。在总体中，X 和 Y 的关系被叙述为

	$Y=1$	$Y=2$
$X=1$	p_{11}	p_{12}
$X=2$	p_{21}	p_{22}

其中，p_{ij} 为 $X=i$ 且 $Y=j$ 的概率，如果我们只有具完整数据的 150 个观察值，似然值为

$$L=(p_{11})^{52}(p_{12})^{21}(p_{21})^{34}(p_{22})^{43},$$

受到限制，此 4 个概率必须加总为 1。此 4 个概率的 ML 估计值是每一个单元格中的简单比率，也就是

$$\hat{p}_{ij}=\frac{n_{ij}}{n},$$

其中 n_{ij} 为落入单元格 (i, j) 中的个案数目。因此我们得到

$$\hat{p}_{11}=0.346,$$

$$\hat{p}_{21}=0.227,$$

$$\hat{p}_{12}=0.140,$$

$$\hat{p}_{22}=0.287.$$

但这不可行，因为我们还有额外的仅有 Y 的观察值，需要把它们整合到似然值中。假设缺失数据为可忽略的，有 $Y=1$ 的个案其似然值就是 $p_{11}+p_{21}$，即 $Y=1$ 的边际概率。同样地，对于 $Y=2$ 的个案，似然值为 $p_{12}+p_{22}$。因此，整个样本的似然值为

$$L=(p_{11})^{52}(p_{12})^{21}(p_{21})^{34}(p_{22})^{43}(p_{11}+p_{21})^{19}(p_{12}+p_{22})^{31}.$$

我们如何找到使这个表达式最大化的 p_{ij} 值？对大部分关于缺失数据问题的 ML 运用而言，这些估计值并没有明确的解法。然而，迭代的方法是必要的。但在这个例子中，形态必定是单调的(因为只有一个变量有缺失数据)，因此，我们可以分别估计给定 Y 时 X 的条件式分布及 Y 的边际分布。然后，我们结合所有结果以得到 4 个单元格的概率。对于 2×2 的表，ML 估计量的一般形式为

$$\hat{p}_{ij} = \hat{p}(X = i \mid Y = j)\hat{p}(Y = j)$$

右边的条件式概率仅利用完整数据的个案来估计。它们可以通过一个普遍的方法——即将 2×2 表中单元格频数除以栏总数——而获得。Y 的边际概率估计值可由加总栏频数及缺失值的个案的 Y 频数，再除以样本数而获得。因此，我们得到

$$\hat{p}_{11} = \left(\frac{52}{86}\right)\left(\frac{86+19}{200}\right) = 0.3174$$

$$\hat{p}_{21} = \left(\frac{34}{86}\right)\left(\frac{86+19}{200}\right) = 0.2076$$

$$\hat{p}_{12} = \left(\frac{21}{64}\right)\left(\frac{64+31}{200}\right) = 0.1559$$

$$\hat{p}_{22} = \left(\frac{43}{64}\right)\left(\frac{64+31}{200}\right) = 0.3591$$

当然，这些估计值与只用具完整讯息个案所得到的估计值不相同。另一方面，一个普遍使用的二分类变量相关性的测量的交叉相乘比，不管是从 ML 估计值计算而来还是仅根据完整个案而得的估计值，都会是相同的。简而言之，有 X 缺失数据的观察值没有为我们提供关于交叉相乘比的额外信息。

这个例子用于说明具缺失数据的 ML 估计的一些普遍特征。然而，很少读者会在应用 ML 估计法时，像以上所讲述的用手计算他们的特别运用。我们需要一个可以处理较多数据形式以及缺失数据型态的通用的软件。虽然对于分析列联表的 ML 估计计算并不困难(Fuch, 1982; Schafer, 1997)，事实上也没有商业软件可以处理这个任务。但在网络上可获取免费的软件：

(1) Jeroen K. Vermunt 的 Windows 版 ℓ_{EM}程序，用以估计当某些数据缺失时各种类型的类别数据模型。

(http://www. kub. nl/fac-ulteiten/fsw/organisatie/departmenten/mto/software2. htlm)

(2) Joseph Schafer 的 CAT 程序，用以估计缺失数据的分层对数线性模型，但目前只在 S-PLUS 软件包中作为一个程序库可供利用。

(http://www. stat. psu. edu/~jls)

(3) David Duffy 的 LOGLIN 程序，用以估计缺失数据的多种对数线性模型(http://www2. qimr. edu. au/davidD)。

具正态分布数据的线性模型

在数据来自多元变量的正态分布的假设下，ML 可以用来估计许多线性模型。可能的模型包括普通线性模型、因子分析、联立方程和具潜变量的结构方程。虽然多元变量正态性的假设很强，但对于没有缺失数据的变量是全然无害的。此外，甚至当有些变量具非正态分布的缺失数据时(如虚拟变量)，多元变量正态性假设下的 ML 估计值通常有好的特性，尤其是当数据为 MCAR 时。[6]

有许多方法可以用来对具可忽略缺失数据机制的多元变量正态数据做 ML 估计。当缺失数据服从一单调形态时，可用先前叙述过的方法将似然因子化运用到用传统软件估计的条件式及边际分布中(Marini, Olsen & Rubin 1979)。然而，这个方法因可能的运用而受到较多限制，而且不容易得到好的标准误及检验统计量的估计值。

一般缺失数据模式可用一个称为期望最大化的算法来处理(Dempster, Laird & Rubin, 1977)，它可产生平均数、标准差(或同样地可产生平均数和协方差矩阵)和相关性的 ML 估计值。这些总和统计量接着可被输入到标准线性模型化软件中以得到所关注参数的一致估计值。EM 方法的优点有：(1)它很容易使用；(2)很多商业的或免费的软件都可以执行。同时也有两个缺点：(1)由线性模型化所报告的软件标准误和检验统计量并不正

确；(2)对于过度识别模型，估计值不是全然有效的（“过度识别模型”指那些在协方差矩阵上隐含限制的模型）。

一个比较好的方法是直接最大化所假设的线性模型的多元变量正态似然值。直接的 ML(有时也称为原始最大似然)提供具正确标准误的有效估计值，但需要有难以掌握的专业的软件。在本章接下来的部分，我们会具体讲述如何使用 EM 算法和直接 ML。

EM 算法

当有些数据缺失的时候，EM 算法是取得 ML 估计量的一个非常普遍的方法(Dempster et al.，1977；McKachlan & Krishnan，1997)。它之所以被称为 EM，是因为它包含两个步骤：一个期望步骤，一个是最大化步骤。这两个步骤在一个迭代的过程中多次重复，最终收敛到 ML 估计值。

在此我并不解释一般设定下 EM 算法的两个步骤，而是着重于它对于多变量正态分布的运用。这里步骤 E 即期望步骤实质上变成缺失值的回归插补。假设数据组有 4 个变量，X_1 到 X_4，且每个变量都含有不具特殊形态的缺失值。我们从选择未知参数的起始值，也就是平均数和协方差矩阵开始。这些起始值可以通过样本平均数和协方差的标准公式获得，不管是用成列删除或成对删除。根据参数的起始值，我们可以得到回归任何 1 个 X 于其他 3 个 X 上的系数。例如，假设有一些个案 X_1 和 X_2 都有数据，但 X_3 和 X_4 没有数据。我们用协方差矩阵的起始值以得到回归 X_3 于 X_1 和 X_2 及回归 X_4 于 X_1 和 X_2 的结果。我们接着用这些回归系数，根据 X_1 和 X_2 的观察值，产生 X_3 和 X_4 的插补值。对于只有 1 个变量数据缺失的个案，我们根据所有其他 3 个变量进行回归插补。

在所有缺失数据都插补完之后，步骤 M 即最大化步骤包括使用插补数据和没有缺失的数据，来计算新的平均数和协方差矩阵的值。对于平均数，我们只要使用一般公式即可。对于方差和协方差，必须使用修正过的公式给每个涉及缺失数据的公式项。明确地讲，必须根据插补过程中的回归方程，对应残差方

差和残差协方差从而增加公式项。例如，假设对观察值 i，X_3 为使用 X_1 和 X_2 所插补。那么，只要$(x_{i3})^2$ 被用于传统的方差公式中，我们就用 $(x_{i3})^2+s_{3\cdot21}^2$ 代替，其中，$s_{3\cdot21}^2$为从回归 X_3 于 X_1 和 X_2 而来的残差方差。加入残差项可以矫正通常在更传统插补方案中产生的对于方差的低估。假设对于观察值 i，X_4 也是缺失的。那么当计算 X_3 和 X_4 的协方差时，只要 $x_{i3}x_{i4}$被用于传统协方差公式中，我们就用 $x_{i3}x_{i4}+s_{34\cdot21}$ 代替。最后一项为 X_3 和 X_4 在控制 X_1 和 X_2 后的残差协方差。

一旦我们得到新的平均数和协方差矩阵的估计值，我们就重新开始步骤 E。也就是说，我们用新的估计值来产生对于缺失值的新的回归插补。我们一直循环步骤 E 和步骤 M 直到估计量收敛，即从一个迭代到另一个迭代之间结果已经几乎不变。

注意 EM 运算法避免了传统回归插补的一个难题——决定使用哪些变量作为自变量，并处理不同数据缺失模式有不同的自变量的情况。因为 EM 总是从完整协方差开始，因此它可能得到对于任何组预测量的回归估计值，不论在某个特别的缺失数据形态中可能存在多么少的个案。因此，EM 总是使用所有可得的变量作为预测量以插补缺失数据。

EM 实例

我们使用来自 1994 年美国新闻和世界报导对美国最好大专院校的指南中关于美国 1302 所大专院校的数据。我们考虑下列变量：

GRADRAT　高年级毕业生对四年前就读人数的比率(100%)

CSAT SAT　语言和数学两个科目的平均分数

LENROLL　新生入学人数的自然对数

PRIVATE　1 = 私立；0 = 公立

STUFAC　师生比率(100%)

RMBD　每年食宿总支出(千美元)[7]

ACT　平均 ACT 分数

我们的目标是要估计一个线性回归模型，GRADRAT 是因变量，其他 5 个是自变量。ACT 不会在回归模型中，它被纳入 EM 估计是因为它与 CSAT (CAST 为有大量缺失数据的变量）高度相关，因此可以让我们得到较好的缺失插补值。

表 4.1　利用可得个案而得的大专学院数据的描述性统计

变　量	非缺失个案	平均数	标准差
GRADRAT	1204	60.41	18.89
CSAT	779	967.98	123.58
LENROLL	1297	6.17	1.00
PRIVATE	1302	0.64	0.48
STUFAC	1300	14.89	5.19
RMBRD	783	4.15	1.17
ACT	714	22.12	2.58

表 4.1 列出了每一个变量没有缺失数据的个案数目，以及这些有数据的个案的平均数与标准差。只有 PRIVATE 这个变量的数据是完整的。自变量 GRADRAT 有 8%的大专院校缺失数据。CSAT 和 RMBRD 都缺失 40%左右，而 ACT 则缺失 45%。对除了 ACT 外的所有变量使用成列删除会产生一个只有 455 个个案的样本，明显令人无法接受。然而，为了比较，我们还是把成列删除回归估计值呈现于表 4.2。

表 4.2　使用成列删除预测 GRADRAT 的回归

变　量	系　数	标准误	t 统计量	p 值
截距	−35.028	7.685	−4.56	0.0001
CSAT	0.067	0.006	10.47	0.0001
LENROLL	2.417	0.959	2.52	0.0121
PRIVATE	13.588	1.946	6.98	0.0001
STUFAC	−0.123	0.132	−0.93	0.3513
RMBRD	2.162	0.714	3.03	0.0026

其次,我们使用 EM 运算法来获得平均数、标准差和相关性的估计值。在主要的商业套装软件中,EM 对数法可得于 BMDP、SPSS、SYSTAT 及 SAS。然而,使用 SPSS 和 SYSTAT ,要存盘输入其他线性模型化例行程序,这非常麻烦。对于大专院校数据,我们使用 SAS 的 MI 程序,结果显示于表 4.3 和4.4。如使用其他 EM 软件,这个程序自动实现前面叙述过的所有步骤。

表 4.3 来自 EM 算法的平均数和标准差

变 量	平均数	标准差
GRADRAT	59.86	18.86
CAST	957.88	121.43
LENROLL	6.17	0.997
PRIVATE	0.64	0.48
STUFAC	14.86	5.18
RMBRD	4.07	1.15
ACT	22.22	2.71

表 4.4 来自 EM 算法的相关性

	GRADRAT	CAST	LENROLL	PRIVATE	STUFAC	RMBRD	ACT
GRADRAT	1.000						
CAST	0.591	1.000					
LENROLL	−0.027	0.192	1.000				
PRIVATE	0.398	0.161	−0.619	1.000			
STUFAC	−0.318	−0.315	0.267	−0.368	1.000		
RMBRD	0.478	0.479	−0.016	0.340	−0.282	1.000	
ACT	0.598	0.908	0.174	0.224	−0.293	0.484	1.000

比较表 4.3 和表 4.1 的平均数可见,不出意料,最大的差异出现在所有有着最多缺失数据的变量中:GRADRAT、CSAT、RMBRD 及 ACT。然而,即使对这些变量而言,成列删除和 EM 结果之间的差异也都没有超过 2%。

表 4.5 显示了使用 EM 统计量得到的回归估计值。虽然系数与使用成列删

除的表 4.2 并没有明显差异，但所得的标准误却低很多，从而导致较高的 t 统计量和较低的 p 值。遗憾的是，在这个例子中虽然系数为真实 ML 估计值，但标准误确实太低，因为它们假设所有个案都有完整数据。为了得到正确的标准误估计值，我们将使用随后叙述的直接 ML 方法。[8]

表 4.5　根据 EM 算法预测 GRADRAT 的回归

变　量	系　数	标准误	t 统计量	p 值
截距	−32.395	4.355	−7.44	0.0001
CSAT	0.067	0.004	17.15	0.0001
LENROLL	2.083	0.539	3.86	0.0001
PRIVATE	12.914	1.147	11.26	0.0001
STUFAC	−0.181	0.084	−2.16	0.0312
RMBRD	2.404	0.400	6.01	0.0001

直接 ML

如我们之前看过的，大多数 EM 运算法的软件产生平均数及不受限制的相关性(或协方差)矩阵的估计值。当这些总和统计量被输入其他线性模型化程序时，产生的标准误估计值将会有偏误，而且通常是被低估的。为了做得更好，我们需要直接把所关注模型的似然值最大化。可以使用任一估计包含潜变量的结构方程模型(SEMs)的软件包来完成这个任务。

当只有少量的缺失数据时，可以使用处理多组的任一 SEM 程序来估计线性模型(Allison，1987；Muthen，Kaplan & Hollis，1987)，其中包含 LISTREL 和 EQS。对于具更普遍形态的缺失数据，目前有 4 个程序可以执行线性模型的直接 ML 估计：

Amos　一个 SEM 模型化的商业软件，现在可以作为一个独立的套装或 SPSS 的一个模块来使用。信息参见 http://smallwaters.com。

Mplus　一个独立的商业软件。信息参见 http://statmodel.com。

LINCS　Gauss 的一个商业模块。信息参见 http://www.aptech.com ≠

3party. com。

Mx 一个免费软件，可从 http://views. vcu. edu/mx 下载。

在进行实例讲解前，让我们先回顾一下基本的理论。令 $f(x \mid \mu, \Sigma)$ 为一个被观察向量 x、平均数向量 μ 及协方差矩阵 Σ 的多变量正态密度。如果我们对于一个来自这个多变量正态分布、有着 $i=1, \cdots, n$ 个观察值的样本有完整数据，那么似然函数则为

$$L(\mu, \Sigma) = \prod_i f(x_i \mid \mu, \Sigma)$$

现在假设我们没有完整数据。如果对于个案 i 有一些变量缺失数据，我们令 x_i 为较小的向量，直接将缺失的元素从 x 中去除。令 μ_i 为 μ 的次向量，以排除在 x_i 缺失的对应的元素，且令 Σ_i 为 Σ 的次矩阵，以删除对应于 x 缺失值的列和栏。我们的似然函数则变为

$$L(\mu, \Sigma) = \prod_i f(x_i \mid \mu_i, \Sigma_i).$$

虽然这个函数看起来非常简单，但却比具完整数据的似然函数更加难以处理。然而，这个似然函数可以由传统的 ML 估计方法来最大化。同时，我们可以取此似然函数的对数，对未知参数偏微分，并令结果等于 0。得到的方程可以使用数学算法来求解，如使用产生标准误如同一个相乘组合的 Newton-Raphson 方法。它也可能将一结构加于 μ 和 Σ 上，从而让它们成为一个相对应于某个假设线性模型的、有较小组参数的函数，例如，因子模型组

$$\Sigma = \Lambda\Phi\Lambda' + \Psi,$$

其中 Λ 为因子载荷矩阵，Φ 为潜在因子的协方差矩阵，Ψ 为误差成分的协方差矩阵。估计过程可以产生这些参数的 ML 估计值和标准误估计值。

直接 ML 实例

在此用同时具有使用者图解界面及文字界面两者的 Amos 3.6 来估计大专

院校的回归模型。图解界面允许使用者用变量中的箭头来指明方程式。因为作者不能做实时示范,故而将相同的文字命令显示于图 4.1 中。数据在一个叫 COLLEGE.DAT 的自由格式文字文件中,缺失数据标示为-9。$mtructure 告诉 Amos 要估计指定变量的平均数,这是估计有缺失数据模型的一个必要的部分。$structure 指令指明要估计的方程。紧接在等号后的括号代表要估计一个截距。方程最后面的(1)error 告诉 Amos 要包含一个系数为 1.0 的误差项。最后一行,act<>error,考虑 ACT 和误差项之间的相关性,这是有可能的,因为 ACT 对于 GRADRAT 没有直接效应。Amos 自动考虑 ACT 和回归方程中的其他自变量相关。

```
$Sample size = 1302
$missing = -9
$input variables
   gradrat
   csat
   lenroll
   private
   stufac
   rmbrd
   act
$rawdata
$include = c:\college.dat
$mstructure
   csat
   lenroll
   private
   stufac
   rmbrd
   act
$structure
   gradrat = ( ) + csat + lenroll + private + stufac
            + rmbrd + (1)error
   act<>error
```

图 4.1 Amos 预测 GRADRAT 的回归模型指令

结果显示于表 4.6。与表 4.5 中两步骤 EM 估计值的比较显示,两者得到的系数相同,但 Amos 标准误显著较大,这正是我们所期望的。但相对于我们在

表 4.2 由成列删除所得到的仍然相当小。

表 4.6 以 Amo 使用直接 ML 预测 GRADRAT 的回归

变 量	系 数	标准误	t 统计量	p 值
截距	−32.395	4.863	−6.661	0.000000
CSAT	0.067	0.005	13.949	0.000000
LENROLL	2.083	0.595	3.499	0.000467
PRIVATE	12.914	1.277	10.114	0.000000
STUFAC	−0.181	0.092	−1.968	0.049068
RMBRD	2.404	0.548	4.386	0.000012

结论

对于随机缺失的数据而言，最大似然可以说是一个有效且实用的方法。在这种情况下，对于大样本，ML 估计值应该是最合适的。对于可用如 LISREAL 程序所估计的一般的结构方程模型之线性模型，ML 估计值很容易从许多广泛可用的软件包中获得。类别数据的对数线性模型 ML 估计也有可用的软件，但在这个设定中执行比较间接。ML 方法的一个限制条件为，它需要一个包含所有缺失变量的联合概率的模型。对这个目的而言，多变量正态模型通常是方便的，但对于许多其他的运用可能不太实际。

第 5 章 | 多重插补:基本原理

虽然 ML 代表了处理缺失数据传统方法的一个重要的发展,但它也有其局限性。如我们已经看到的,关于线性模型和对数线性模型的 ML 理论和软件容易获取,但超出这两者之外的理论抑或软件通常比较缺乏。例如,如果你想要估计一个 Cox 比例风险模型或一个有序的logistic 回归模型,你将难以对缺失数据执行 ML 方法。甚至假使你的模型可以用 ML 估计,但你也可能难以找到你特别需要的、专业的诊断或图形输出软件。

值得庆幸的是,有一个备选方法——即多重插补,它有着与 ML 相同的最适特性,但却排除了某些局限性。更明确地,当数据为 MAR 时,正确使用多重插补(MI)会产生一致的、渐近有效且渐近正态的估计值。不同于 ML, MI 几乎可以被任何一种数据及任何一种模型所使用,且分析可以利用未修改的、传统的软件执行。当然,MI 也有它自身的缺点。它的执行可能很麻烦,且也容易出错。这两个问题可以通过使用好的软件做插补来解决。但它最致命的缺点是,每次你使用 MI 时,它都会产生不同的估计值(但愿差异很小)。这可能导致奇怪的情况,即不同研究者使用相同的方法、相同的数据却得到不同的数字。

单一随机插补

MI 不产生一个单一组数目的原因是,故意在插补过程中引入了随机变异。若没有一个随机成分,决定性的插补方法通常对有缺失数据的变量产生低估的方差估值,而且有时候,协方差也一样会被低估。如我们在前一章所讲到的,对于多元变量正态模型的 EM 运算法,可以使用残差方差和协方差估计值来矫正

传统公式，来解决这个低估方差/协方差的问题。然而，一个好的备选方法是随机抽取自每一个插补变量的残差分布，并把这些随机数字加到插补值。另外，传统公式也可以用来计算方差和协方差。

这里举一个简单的例子。假设我们想要估计 X 和 Y 之间的相关性，但50%的个案有 X 缺失数据。我们可以通过对有完整数据的个案回归 X 于 Y，然后再用得到的回归方程产生有缺失 X 的个案的预测值，进而插补缺失的 X 值。作者对一个具 10000 个个案的模拟样本做了这种处理，其中 X 和 Y 抽取自一个标准双变量正态分布中，两者的相关性为 0.30。一半 X 的值被指定为缺失的(完全随机)。利用回归于 Y 的方法来替代缺失值后，X 和 Y 的相关性估计为 0.42。

样本相关性是样本 X 和 Y 的协方差除以它们样本标准差的乘积。为什么利用了回归于 Y 的方法后会高估相关性呢？首先，回归插补方法产生无偏误的协方差估计值。此外，(没有缺失数据的)Y 的标准差被正确估计为1.0，但(包含插补值的)X 的标准差只有 0.74，而其实际的标准差应为1.0。所以造成了相关度的高估。换一个方法来考虑这个问题，5000 个有缺失数据的个案的 X 插补值为 Y 的一个完全线性函数，因此增大两个变量间的相关性。

我们可以通过从 X 的残差分布随机抽取残差并将这些随机数字加到 X 的预测值上，以矫正这个偏误。在这个例子中，(对 Y 回归的)X 的残差分布为一平均数为 0 且标准差为 0.9525 的正态分布(从成列删除的最小二乘回归估计而得)。对于个案 i，令 μ_i 为一来自标准正态分布的随机抽取且令 $\hat{x}_i$ 为从回归 X 于 Y 而得到的预测值。我们修正的插补值则为 $\tilde{x}_i = \hat{x}_i + 0.9525u_i$。对所有有缺失 X 的观察值而言，我们用 $\tilde{x}_i$ 替代，然后再计算相关性。当作者对这个有10000 个个案的模拟样本进行这种处理时，(有修正插补值的)X 和 Y 之间的相关性为 0.316，只比真实值 0.300 高一些而已。

多元随机插补

随机插补可以消除决定性的插补所特有的偏误，但仍存在一个严重的问

题。如果我们把插补的数据(不管是随机或决定性的)当做真实的数据来使用,导致的标准误估计值通常会比较低,而检验统计量会比较高。标准误估计的传统方法不足以说明数据为插补的事实。

对于使用随机插补而言,其解决方案在于不止一次地重复插补过程,从而产生多个完整数据组。因为随机成分,关注的参数估计值在每个插补的数据组中将会只有微小差异。横跨插补的变异可以被用来向上调整标准误。

对于一个有着 10000 个个案的模拟样本,作者重复随机插补过程 8 次,产生的估计值见表 5.1。虽然这些估计值近似反映无偏误的,但因为没有考虑插补[9],所以标准误是向下偏误的。我们通过对这 8 个相关性估计值取平均值,得到一个单一的估计值 0.3125。用以下 3 个步骤来产生改良的标准误估计值:

(1) 估计标准误平方(以得到方差),并对 8 个结果取平均值;

(2) 对这 8 个复制的结果计算相关性估计值的方差;

(3) 加总步骤 1 和步骤 2 的结果(运用步骤 2 中的一个对方差的小矫正),并取平方根。

将此置入一个方程中,令 M 为复制的次数,r_k 为复制 k 中的相关性,s_k 为复制 k 中估计到的标准误。则 $\bar{r}$ 标准误的估计值(相关性估计值的平均数)为

$$S.E.(\bar{r})=\sqrt{\frac{1}{M}\sum_k s_k^2+\left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k(r_k-\bar{r})^2} \qquad [5.1]$$

这个方程可以用于由多重插补所估计的任何参数,r_k 表示所关注参数的第 k 个估计值(Rubin, 1987)。把这个公式运用于我们关注的例子中,得到标准误为 0.01123,大约较由 8 个样本的标准误的平均数高出 24%。

表 5.1 随机插补数据之相关性与标准误

相关性	S.E.	相关性	S.E.
0.3159	0.00900	0.3118	0.00903
0.3108	0.00903	0.3022	0.00909
0.3135	0.00902	0.3189	0.00898
0.3210	0.00897	0.3059	0.00906

在参数估计值中考虑随机变异

虽然作者之前描述过的插补缺失数据的方法已经相当完善了，但它并不是最理想的。为了产生 X 的插补值，作者对有完整数据的个案回归 X 于 Y 上，以产生回归方程

$$\hat{x}_i = a + by_i$$

对于 X 有缺失数据的个案，插补值计算为

$$\tilde{x}_i = a + by_i + s_{x\cdot y}u_i$$

其中 u_i 是一个来自标准正态分布的随机抽取，而 $s_{x\cdot y}$ 为误差项的估计标准均方根差。对于模拟的数据组，我们得到 $s_{x\cdot y} = 0.9525$。这些值用来对 8 个完整数据组中的每一个产生插补值。

这个方法的问题在于，它视 a、b 和 $s_{x\cdot y}$ 如同真实参数，而非样本估计值。很明显，我们不能得知真实值为多少，但对于"适当"的多重插补(Rubin，1987)，每一个插补的数据组应该基于不同组的 a、b 和 $s_{x\cdot y}$ 值。这些应该从参数的贝叶斯后验分布中随机抽取。只有这样，多重插补才能完全体现我们关于未知参数的不确定性。

这种观点自然会产生许多疑问。什么是参数的贝叶斯后验分布？我们如何从后验分布中随机抽取用于插补的值？我们真的需要这额外的复杂程序吗？第一个问题需要另一本书来回答，而且幸好在社会科学的量化运用系列丛书里(Iversen，1985)有一本相关的书。对于第二个问题，有数个不同的方法可以从后验分布中进行随机抽取，它们有些被包含在容易使用的软件里。在本章稍后部分，当我们考虑在多变量正态模型下的 MI 时，作者将解释一种叫做数据扩增法的方法(Schafer，1997)。

如果不使用从后验分布中随机抽取的方法可以吗？这个问题的答案很重要，因为有一些随机插补软件，如 SPSS 中的缺失数据模块，可以随机抽取参数值。在许多例子中，作者认为答案是可以的。如果样本够大且有缺失数据的个

案比例很小，那么缺少这个额外步骤的 MI 就更容易产生非常接近那些包含这个额外步骤的结果。另一方面，如果样本小或有缺失数据的个案比例很大，则额外的变异可以产生明显的差异。

继续我们所关注的例子。作者用数据扩增法插补 8 个新数据组，以从参数的后验分布中产生随机抽取。表 5.2 提供 X 和 Y 间的相关性及每一个数据组的标准误。相关性估计值的平均数为 0.31288。使用方程 5.1，估计标准误为 0.01329，比由较粗糙的插补方法所获得的 0.01123 稍大。一般而言，当用于插补中的参数为随机抽取时，标准误会略显较大。

表 5.2　使用数据扩增法的随机插补数据之相关性与标准误

相关性	S.E	相关性	S.E
0.30636	0.0090614	0.32086	0.0089705
0.31316	0.0090193	0.29760	0.0091143
0.31837	0.0089864	0.32701	0.0089306
0.31142	0.0090302	0.30826	0.0090498

在多变量正态模型下的多重插补

为了做多重插补，你需要一个模型以产生插补值。对于刚才考虑过的两个变量的例子，作者利用一个具正态分布误差的简单回归模型。很明显，更复杂的情况需要更复杂的模型。然而，对于模型的选择来说，MI 可能没有 ML 那么敏感，因为模型只用来插补缺失数据，而不用来估计其他参数。

理想的情况是，我们会特别建构插补模型以代表每一个数据组的独特的特征。在实际应用中，利用便于使用的、且适用于广泛数据组的、合理良好近似反映的、现成的模型会更加方便。

MI 最受欢迎的模型为多变量正态模型，也就是先前在第 4 章使用过的、具缺失数据的线性模型的 ML 估计的基础。多变量正态模型意味着：

(1) 所有变量都有着正态分布；

(2) 每一个变量可以用所有其他变量的一个线性函数以及一个正态、同方差的误差项所表示。

虽然这些是很严格的条件，但实际上，即使当有些变量已经明确不是正态分布时，多变量正态模型仍然可以实现好的插补(Schafer, 1997)。对那些没有缺失数据的变量而言，这是一个完全无害的假设。而对那些有缺失数据的变量而言，正态化转换则可以大大地改善插补的质量。

本质上讲，多变量正态模型下的 MI 是前述两变量例子中使用过的方法的概括。对于每一个有缺失数据的变量，我们估计该变量对于所有其他受关注变量的线性回归。最理想的结果是，回归参数是从贝叶斯后验分布中随机抽取而得。估计到的回归方程接着用来产生有缺失数据的个案的预测值。最后，对每一个预测值，我们加上该变量残差正态分布的一随机抽取值。

插补过程中最复杂的部分是从贝叶斯后验分布中得到随机抽取值。在作者写作本书之时，已有两种用以实现随机抽取的算法可用现有的软件执行：数据扩增法(Schafer, 1997)及抽样重要性/重抽样(SIR; Rubin, 1987)。这里给出一些执行这些方法的计算机程序：

数据扩增法

NORM　由 Schafer 开发的免费软件，并在其 1997 年出版的书中叙述过。作为一个独立的 Windows 版本或 S-PLUS 的一个程序库，相关资料参见 http://www.stat.psu.edu/~jls/。

SOLAS　一个独立的商业套装，包含数据扩增法(版本 2 及更高版本)和一个倾向评分方法。后者在许多应用上是无效的(Allison, 2000)。相关资料参见 http://www.statsolusa.com。

PROC MI　一个 SAS 程序，可于 8.1 或更高的版本中使用，相关资料参见 http://www.sas.com。

抽样重要性/重抽样

AMELIA　一个由 King, Honaker, Joseph, Scheve 及 Singh(1999)开发的免费套装。作为一个独立的 Windows 程序或 Gauss 的一个模块使用。相关资料参见 http://gKing.harvard.edu/stat.shtml。

SIRNORM　一个由 C. H. Brown 和 X. Ling 写的 SAS 宏命令。相关资料参见 http://yates. coph. usf. edu/research/psmg/web. html。

两种算法理论上都有一定的合理性。SIR 的提倡者(King, Honaker, Joseph & Scheve, 2001)宣称它需要极少的计算机运算时间。然而,这两种方法的相对优越性尚无定论。因为作者有较多关于数据扩增法的经验,因此将会在本章接下来的部分着重介绍这个方法。

多变量正态模型的数据扩增法

数据扩增法是马尔可夫链蒙特卡尔运算法的一种形式,一个寻找后验分布的普遍方法,在贝叶斯统计学中越来越受欢迎。在这个部分,作者将叙述它如何处理多变量正态模型。虽然现成可用的软件可自动执行大部分的运算,但对于实际进行过程而言,有一个普遍的理解是有帮助的,特别是当事情出错时。

迭代算法的普遍结构与在前一章叙述过的多变量正态模型的 EM 算法很像,但它还需在两个时点做随机抽取,接下来作者会叙述这点差异。在开始 DA 前,必须选择用于插补过程的一组变量。这组变量很明显应该包含所有有缺失值的变量,以及其他在模型中需要进行估计的变量。同时也应该把不在目标模型中但与具缺失数据的变量高度相关的、或与那些具缺失数据的变量的概率有关的、额外的变量包含进来。

一旦选定变量后,DA 包含下列步骤:

(1) 选择参数的起始值。对于多变样正态模型,参数为平均数和协方差矩阵。起始值可以利用成列删除或成对删除的标准公式得到。有在前一章叙述过的用 EM 算法而获得的估计值则更好。

(2) 用平均数和协方差的当前值来获得方程回归系数的估计值,方程中每一个有缺失数据的变量回归于所有被观察到的其他变量上。对每一种缺失数据的形态都这样处理。

(3) 用回归估计值产生关于所有缺失值的预测值。对每一个预测值,加上一个从该变量残差正态分布中得到的随机抽取值。

(4) 用具观察值和插补值的"完整"数据，利用标准公式重新计算平均数和协方差矩阵。

(5) 从新得到的平均数和协方差的后验分布中随机抽取平均数和协方差。

(6) 用随机抽取的平均数和协方差，回到步骤 2 且继续循环其后的步骤，直到达成收敛。用在最后一次迭代中所产生的插补值构成一个完整数据组。

步骤 5 需要进一步的解释。为了得到参数的后验分布，我们首先需要一个先验分布。虽然可以根据关于这些参数先前的看法，但是通常的做法是使用一个"不提供信息"的先验分布，即包含很少或根本不包含与参数相关的信息的先验分布。我们看看它如何在一个简单的状态下起作用。假设我们有测量某个单一正态分布的变量 Y 的 n 个样本数。样本平均数为 $\bar{y}$，样本方差为 s^2。我们想要从 μ 和 σ^2 的后验分布中随机抽取平均数和方差。利用一个不提供信息的先验分布，[10]我们可以从自由度为 $n-1$ 的卡方分布中进行抽样、对抽取值取导数、并将结果乘以 ns^2，以得到方差的随机抽取值 $\tilde{\sigma}^2$。接着我们从一平均数为 $\bar{y}$ 且方差为 $\tilde{\sigma}^2/n$ 的正态分布中抽样，以得到平均数的一个随机抽取值。

如果没有缺失数据，这些会是来自真实参数的后验分布的随机抽取值，但如果我们有插补缺失数据，那么实际上有的值是当插补数据为真实数据时所得到的后验分布的随机抽取值。同样，在给定当前参数值的条件下，当我们随机在步骤 3 中插补缺失数据时，有的值为缺失数据的后验分布的随机抽取值。然而，因为当前值可能不是真实值，插补数据也可能不是真实后验分布的随机抽取值，这也就是为什么程序必须是迭代的。通过持续地在参数的随机抽取（取决于观察到的及插补数据）和缺失数据的随机抽取（取决于当前参数）之间不断来回，我们最终得到从数据和参数两者的联合后验分布中得到的随机抽取值，而这个仅取决于被观察到的数据。

在数据扩增法中收敛

当你运用数据扩增法时，你必须指明反复的次数。然而，这产生了一个难

题:需要多少次的迭代才能得到缺失数据和参数的联合后验分布的收敛值?如使用在 EM 运算法中的最大似然迭代估计,估计值会收敛到单一组的值。接着收敛可以很容易地通过检查从一个迭代到另一个迭代参数估计值改变的大小来进行评估。另一方面,对于数据扩增法,运算会收敛到一个概率分布,而非单一组的值。很难确定收敛事实上是否达成。虽然有些诊断统计量可以用来评估收敛(Schafer, 1997),但它们根本不可靠。

在大多数运用中,迭代次数的选择都是胡乱瞎猜。关于其可能的范围,为了给出一些提示,Schafer(1997)在他书里的例子中,使用介于 50 到 100 之间的迭代次数。次数越多越好,但每一次迭代其运算强度可能很大,特别是对有许多变量的大样本来说。指明一个较大的迭代次数可能会让你长时间痛苦地盯着你的屏幕。

有许多原则要谨记在心。第一,缺失数据(事实上是缺失信息,与缺失数据不大相同)的比例越高,就需要越多次的迭代以达成收敛。如果只有 5%的个案有缺失数据,你用小数目的迭代次数就可能过得去。第二,EM 运算法的收敛速度是数据扩增法收敛速度的有效暗示和征兆。一个好的经验法则是,DA 的迭代次数至少要与 EM 所需的迭代次数相同。这也是为什么总在数据扩增法之前进行 EM 的另一个原因(第一个原因是 EM 为数据扩增法提供了良好的起始值)。

作者对于迭代法这个议题的感触是,在大多数的运用中它不是那么重要。从决定性的插补到随机的插补是一个巨大的改进,甚至就算这个随机插补其参数并不是随机抽取的,也已经是巨大的改进。而从没有随机抽取参数的随机插补到有随机抽取参数的随机插补是另一个大的改进,但这个改进毫不起眼。从迭代次数少的数据扩增法到迭代次数多的是更进一步的改进,但在大多数运用中,其边际回报可能相当小。

另一个复杂性来源于多重插补产生多元数据组的事实。至少需要两个数据组,越多越好。给定固定的运算时间,我们可以产生更多数据组,或对每一个数据组产生更多次数据扩增法的迭代。可惜的是,有着很多缺失信息的数据组同时需要更多的迭代和更多数据组。虽然有关这个议题写得很少,但作者更倾

向于优先考虑额外的数据组。

连续的数据扩增法相对平行的数据扩增法

我们刚才已经看到如何使用数据扩增法来产生单一完整的数据组。对于多重插补，我们需要数个数据组。有两个方法被提出用来执行此任务：

(1) 平行的。对每一个想要的数据组进行一个别系列的迭代。这可以从同一组起始值开始(如 EM 估计值)，也可以从不同的起始值开始。

(2) 连续的。进行一长系列的数据扩增法循环。取每第 k 次迭代产生的插补，k 为原来给定数据组的想要的数目。例如，如果我们有 5 个数据组，我们可以先进行 500 次迭代，再接着使用产生自每第 100 次迭代的插补。在第一次插补前，500 次大数目的迭代形成了一个允许该过程收敛到正确分布的测试时期。

两种方法都可以接受。连续方法的一个优点是，较容易收敛到真实后验分布，特别是对于那些在序列中位置较后的数据组。然而，每当你从相同的一系列迭代中取出多个数据组时，却不能确定那些数据组是否在统计上相互独立，而这种独立性是有效推论的必需条件。在相同系列中 2 个数据组越接近，它们就越有可能存在某些依赖性/相关性。这也就是为什么你不能只运行 200 次迭代以得到收敛，并且接着用 5 次迭代来产生 5 个数据组。

平行方法避免了依赖性/相关性的问题，但更不能确定是否达到收敛。此外，Rubin(1987)和 Schafer(1997)都建议与其对每一个序列都使用相同组的起始值，不如从一个以 EM 估计值为中心的“过度分散的”先验分布中抽取起始值，但这并不总是容易执行。[11]

对于大量广泛的运用来说，作者认为选择连续或平行方法不会有重大差异。若有着同样次数的迭代，两种方法应该会给出几乎相同的结果。作者相信在大多数的例子中，当使用平行方法时，以 EM 估计值为每一个迭代系列的起始值，都能获得可接受的结果。

对非正态或类别数据使用正态模型

可以以多变量正态分布紧密近似反映的数据组事实上很稀有，有时会出现具高度偏态分布的变量以及其他完全是类别型的变量。在诸如此类的例子中，我们刚才考虑过的根据正态的方法有任何的价值吗？如稍早前叙述过的，对于没有缺失数据的变量都没问题，因为它们没有缺失数据的变量而且不需要被插补。

对于有缺失数据的变量，大量的证据显示这些插补方法可以处理得相当好，甚至当分布明显不是正态的时候(Schafer，1997)。然而，有一些技巧可以改善插补非正态变量的正态模型的绩效。

对于高度偏态量化变量，在执行插补前先转换这些变量以降低偏度，通常是有帮助的。任何可以胜任这项任务的转换应该都是可行的。在数据已经被插补后，相反的转换可以运用来将该变量变回其原始的度量标准状态。例如，对数转换能大量降低大多数收入数据的偏度；在插补后，只要取收入的反对数即可。这对于有限制范围的变量帮助特别大。如果你插补收入的对数而非收入本身，就不可能产生小于 0 的收入插补值。同样地，如果要插补的变量是比例，logit 转换就可以防止大于 1 或小于 0 的插补。

有些软件可以用另一种方法处理有限制范围的问题。如果你对一个特别变量指明一个最大或最小值，它将会拒绝所有在这个范围外的随机抽取值，并简单地做额外的抽取直到它处于指定范围中。虽然这是一个很有用的可选方法，但使用转换以降低变量中的偏度仍是令人满意的。

对于离散的量化变量，通常需要把插补值做合适的四舍五入以变成一个离散的量尺。例如，假设成人被问及他们有多少个小孩。对这个问题的回答，其分布会是典型的偏态，所以可以通过运用对数或从平方根转换开始。在插补后，反向转换将会产生非整数的值。这些可以被四舍五入取整数以符合原始量尺。有些软件可以自动执行诸如此类的四舍五入。

若完全是类别的变量又怎样呢？虽然有方法和计算机程序设计仅用于只

有类别变量的数据组，以及有类别的和正态分布变量的混合的数据组，但这些方法更加难以使用且通常会彻底地失效。许多使用者也会运用有较少修改的正态模型来执行。二分类的变量，如性别，通常使用有着 0 或 1 的虚拟(指标)变量来代表。任何对一个二分类变量的转换将仍产生一个二分法，因此没有值可以用来试着降低偏度。相反，我们可以如同其任何其他变量一样简单地插补该 0－1 变量。接着根据该插补值是否高于或低于 0.5，再四舍五入插补值至 0 或 1。大部分的插补会落于(0，1)这个区间内，有时候仍会落在该区间外。在这个例子中没有问题，因为我们根据插补值较接近 0 或 1 者，来指派 0 或 1 的值。

如果有超过两个类别的变量，通常由虚拟变量组来表示。在数据扩增法阶段不需要做任何特别的事，但当指派最终值时则需要小心。问题是我们需要指派个体至一个类别且只有这么一个类别，并适当编码所有的虚拟变量。假设被插补变量为婚姻状态，有三个类别：从未结婚、目前已婚、曾结过婚。令 N 为从未结婚的虚拟变量，令 M 为目前已婚的虚拟变量。用这两个变量做插补且使用插补值产生最终的编码。这里有一些可能的插补和得到的编码：

插补值			最终值	
N	M	$1-N-M$	N	M
0.7	0.2	0.1	1	0
0.3	0.5	0.2	0	1
0.2	0.2	0.6	0	0
0.6	0.8	−0.4	0	1
−0.2	0.2	1	0	0

基本原则是这样的。除了 2 个插补值外，也要计算 1 减去这 2 个插补值的总和，这可被视为参照组的插补值。接着确定哪一个类别有最高的插补值。如果该值对应着一个明确为虚拟变量的类别，则指派 1 给该变量。如果最高的值对应参照组，则指派 0 给另外两个虚拟变量。此外在这个背景下负值可能显得比较奇怪，但此法仍然可以运用。延伸至 4 个或更多个类别应该就很容易理解且简单易做了。

探索分析

很多数据分析包含探索工作，在其中分析实验的各种方法和模型。对于任何已经做过这种工作的人而言，多重插补的过程似乎问题比较大。对数个数据组同时执行探索分析必然是一个烦琐的过程。此外，对每个数据组的分析也可能建议使用有细微差异的模型，但多重插补对所有数据组要求一个相同的模型。

解决方法很简单，但却显得特别随意。当产生多元数据组时，只要产生比你做多重插补分析时所需要的多 1 个的数据组即可。因此，如果你想要做 3 个数据组的多重插补分析，就会产生 4 个数据组。之后再用这个额外的数据组做探索分析。一旦你决定某单一模型或小组模型，就要对剩余的数据组重新估计这些模型，并运用我们已经讨论过的结合结果的方法。要谨记，虽然从探索分析获得的参数估计值将近似反映无偏误的，但是所有的标准误会向下偏误且统计检定量会向上偏误。因此，使用更加保守的准则而非通常（有着完整数据）的准则来衡量一个给定模型的适当性可能结果会更令人满意。

MI 实例 1

我们现在有足够的背景知识可以考虑一个现实中的多重插补的例子。让我们重新回顾一下第 4 章中使用过的例子，一个包含 7 个变量测量的 1302 所美国大专院校数据组，其中除了一个变量外，所有的变量都有缺失数据。和之前一样，我们的目标是估计一个预测 GRADRAT 的线性模型，GRADRAT 为高年级毕业生对于四年前以新生身份就读的数目之比率。自变量包含除了 ACT（即 ACT 分数之平均数）以外的所有变量。这个变量（ACT）被包含于插补过程中以得到较佳的 CSAT（即结合 SAT 分数的平均数）预测。后面的变量（CSAT）40%的个案有缺失数据，但对于同时有两个变量（CSAT 和 ACT）数据呈现的 488 个个案，CSAT 与 ACT 高度相关（$r = 0.91$）。

第一个步骤要检查变量的分布以检验正态性。直方图和正态概率图显示除了一个变量外,所有变量都合理接近正态分布。这个例外是高度左偏的就学这个变量。如同在 ML 实例中,作者使用就学的自然对数,其分布有着很小的偏度。

为了执行数据扩增法,作者使用 SAS 的 PROC MI。第一步是要用 EM 运算法来估计平均数、标准差和相关性,结果已在表 4.4 中呈现。EM 运算法用了 32 次迭代达到收敛。这是一个不算太大的数目,其或许反映了某些变量有大比例的缺失数据。然而,它还没有大到意味着运用 EM 运算法或数据扩增法会有严重问题。

这里有一个最小组的 SAS 命令语句以产生多重插补:

```
proc mi data = college out = collimp;
    var gradrat csat lenroll private stufac rmbrd act;
run;
```

college 为输入数据组(对于缺失值用点表示)的名称,而 collimp 为输出数据组(包含被观察到的值与插补值)的名称。var 述句给予用于插补过程中的变量的名称。PROC MI 的默认值为根据一系列连续性的迭代产生的 5 个完整数据组。以 EM 估计值作为起始值,在第一次插补前有 200 次"测试"迭代。接着有 100 次在连续的插补之间的迭代。5 个数据组被写入一个大的 SAS 数据组以利于后面的分析。输出数据组包含一个新的变量_imputation_,其值由 1 到 5 表示不同的数据组。因此,原始数据组有 1302 个观察值,新数据组有 6510 个观察值。

不采用默认值,作者实际上采用了一个稍微复杂的程序:

```
proc mi data = my.college out = collimp seed = 1401;
    minimum = 0600..0126011
    maximum = 1001410...100870031
    round = 1111.111;
    var gradrat csat lenroll private stufac rmbrd act;
    MCMC nbiter = 500 niter = 200;
run;
```

seed = 1401 对于随机数字产生器设定一个种子值，以至于结果可以在后面一次执行中确实能被重新产生。maximum 和 minimum 选择设定对每一个变量的最大和最小值。如果一个随机插补值恰巧落在这些边界外，则该值会被拒绝，而一个新的值会被抽取。对 gradrat 和 stufac，最大值和最小值的理论边界为 0 和 100。对于 lenroll 和 private，没有指明边界。对于 csat、rmbrd 和 act，作者使用每一个变量观察到的最大值和最小值。round 选择四舍五入所有变量的插补值至整数，除了 STUTAC 以外，它被四舍五入到小数点后第一位。

MCMC 述句允许对数据扩增法过程采取更多控制。作者在此已指明在第一次插补前要有 500 次测试的迭代，接着在连续的插补之间有 200 次迭代。

这些迭代能够足以达到收敛吗？答案很难确定，但我们可以试验由 Schafer (1997)建议的一些收敛诊断。一个简单的做法是检查产生自每一个迭代的一些参数值，并看所有迭代得到的结果之间是否有任何趋势。对于有 7 个变量的多元正态模型，参数为 7 个平均数、7 个方差和 21 个协方差。不检查所有参数，只专注于那些牵涉有最多缺失数据的变量的参数，因为这些牵涉有最多缺失数据的变量的参数最有可能出现问题。对于这些数据，变量 CSAT 有 40%的缺失数据，因为终极目标是要估计预测 GRADRAT 的回归，因此我们看一下 CSAT 的二变量回归斜率，也就是 CSAT 和 GRADRAT 的协方差除以 CSAT 的方差。图 5.1 画出了数据扩增法前 100 次迭代的回归斜率值。在第一次迭代后，在斜率系数估计值中似乎没有特别的趋势，这也让人比较放心。

另一个被提议使用的诊断方法，是对所关注的参数于连续迭代中的许多滞后值做一组自相关。目标是要在插补间有足够的迭代，从而使自相关为 0。使用全部系列 1300 个迭代，图 5.2 画出了具不同滞后值的二变量回归斜率值间的自相关，因此最左边的值 0.34 代表距离为一个迭代远的参数值间的相关性。第二个值为两个迭代分开的参数值间的相关性。虽然这两个初始值比较高，但自相关迅速降低至相当低的值(落在 0 的 0.10 间)，明显有一个随机的形态。这两个诊断结合一起，建议我们在插补间可以使用远少于 200 次的迭代。多做一些迭代没有什么不利，而且这里使用的诊断不保证会达到收敛。

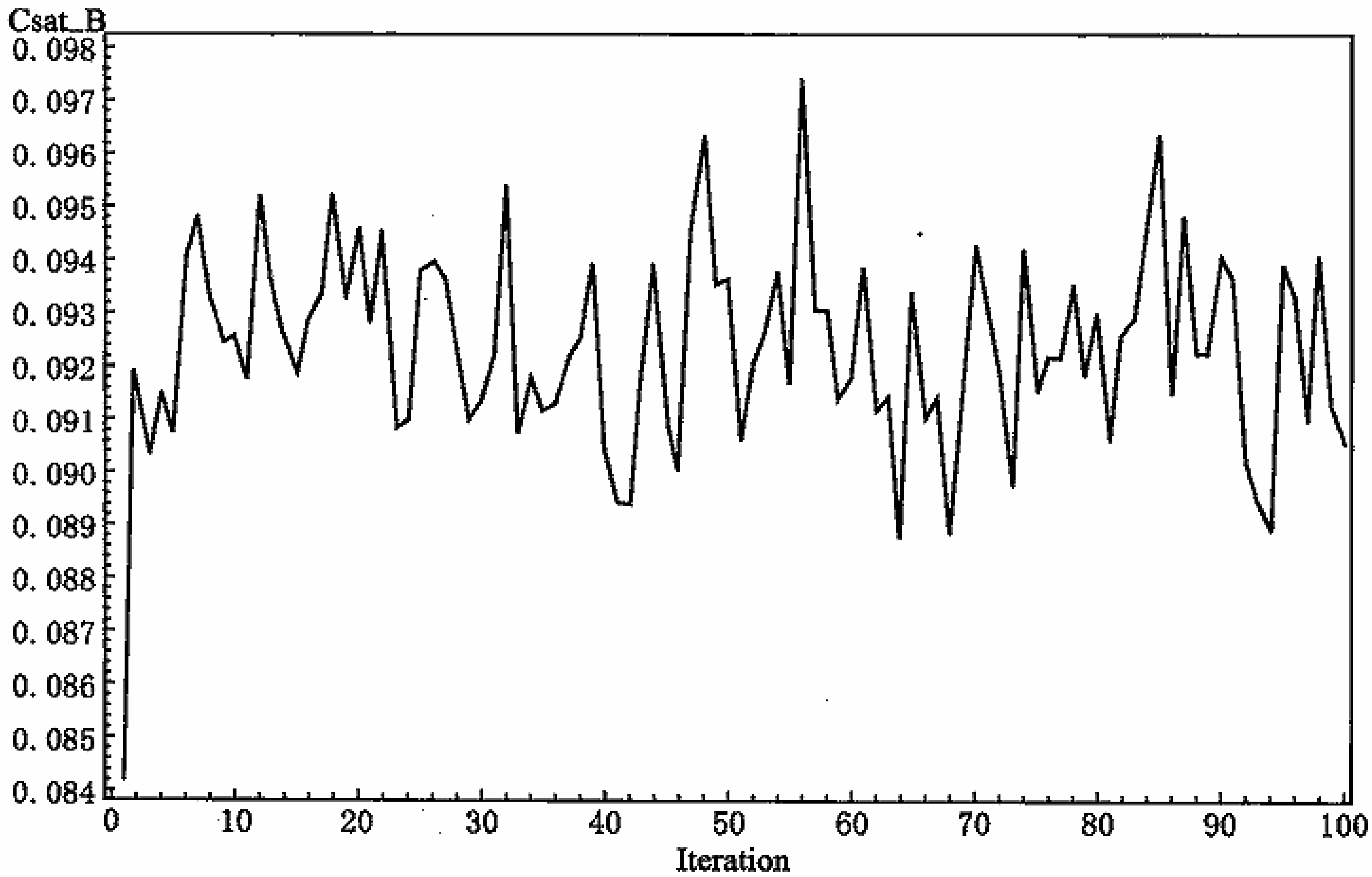


图 5.1 前 100 数据扩增迭代的回归中变量 CSAT(对 GRADRAT)的斜率估计值分布情况

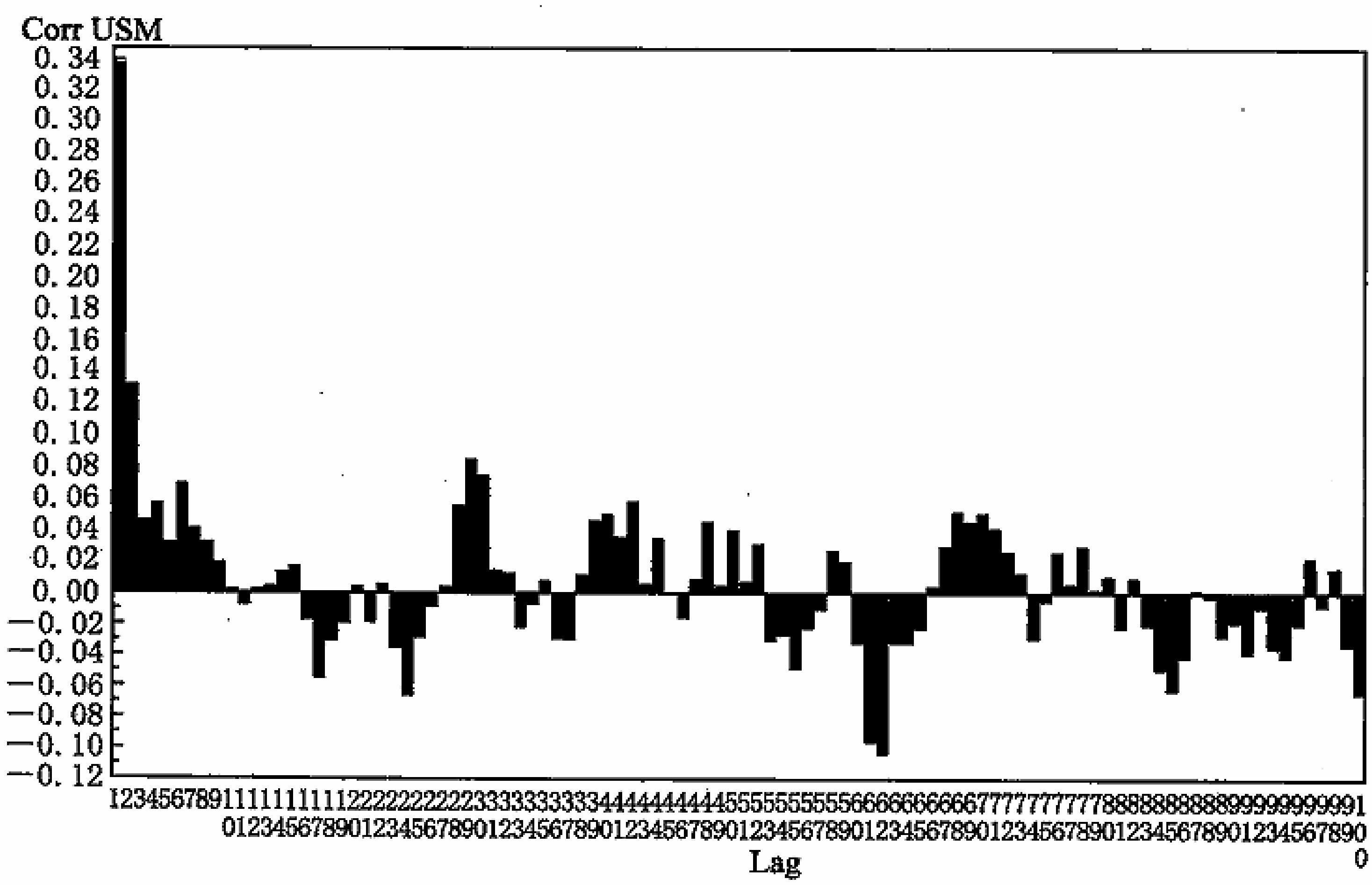


图 5.2 CSAT 对 GRADRAT 的回归斜率的第 1 至 100 个滞后值间变化之自相关

在产生完整数据组后，作者可以将转换过的变量就读(ENROLLMENT)的对数转换回其原始形式，但因为作者期望就读(ENROLLMENT)对于毕业率的效应有逐渐降低的回报(RETRUNS)，因此作者决定让这个变量保留对数形式，就如同作者在第4章对以ML估计的回归模型一样。

表5.3　5个完整数据组的回归系数(及标准误)

截　距	−33.219	(4.272)	−33.230	4.250	−31.256	(4.306)	−34.727	(4.869)	−29.117	(4.924)
CSAT	0.069	(0.004)	0.067	(0.004)	0.071	(0.004)	0.068	(0.004)	0.065	(0.004)
LENROLL	1.550	(0.534)	2.023	(0.526)	1.852	(0.546)	2.187	(0.532)	1.971	(0.538)
PRIVATE	11.632	(1.124)	12.840	(1.126)	12.274	(1.157)	13.468	(1.121)	12.191	(1.141)
STUFAC	−0.145	(0.083)	−0.116	(0.082)	−0.213	(0.084)	−0.142	(0.083)	−0.231	(0.084)
RMBRD	2.951	(0.390)	2.417	(0.392)	1.657	(0.408)	2.103	(0.391)	2.612	(0.393)

所以下一个步骤就是简单地对5个完整数据组中的每一个估计回归模型。通过使用SAS中的BY述句可以使执行变得更容易，从而避免指明5个不同的回归模型：

```
proc reg data = college outset = estimate covout;
    model gradrat = csat lenroll private stufac rmbrd;
    by_imputation_;
run;
```

这组述句告诉SAS，要对由5个_imputation_值所定义的每一个次群体分别估计一个回归模型。outset = estimate要求回归估计值被写入一个叫做estimate的新数据组，而covout要求回归参数的协方差矩阵被包含进该数据组中。这使得在下一步骤中结合估计值显得容易。5个回归结果显示于表5.3中。很显然，每一个回归到另一个回归具有很大的稳定性，但也有值得注意的变异，归因于插补的随机成分。这些回归的结果用另一个叫做MIANALYZE的SAS程序，整合成一个单一估计值组。由下列述句引起：

```
proc mianalyze data = estimate;
    var intercep csat lenroll private stufac rmbrd;
run;
```

这个程序直接对包含有回归运作所产生的系数和相关统计量的数据组estimate来起作用。结果呈现于图 5.3。

Variable	Mean	Std Error Mean	DF	t for H0: Mean=0	Pr>|t|	Fraction Missing Information
intercept	−32.309795	5.639411	72	−6.596995	<0.0001	0.255724
csat	0.068255	0.004692	39	14.547388	<0.0001	0.356451
lenroll	1.916654	0.595229	110	3.220027	0.0017	0.206210
private	12.481050	1.367858	40	9.124524	<0.0001	0.344151
stufac	−0.169484	0.099331	42	−1.706258	0.0953	0.329284
rmbrd	2.348136	0.670105	10	3.504132	0.0067	0.708476

Multiple-Imputation Parameter Estimates

图 5.3　从 PROC MIANALYZE 得出的部分结果

图 5.3 中标示为“平均数”的字段包含表 5.3 中系数的平均数。使用方程 5.1 计算的标准误明显比表 5.3 中的标准误大，因为不同回归间的变化被加到回归内的变化了。然而，对于一些系数而言，有比其他系数更多的回归间的变化。在低端，lenroll 系数的标准误在图 5.3 中只比表 5.3 的标准误平均值大了约 10%。在高端，rmbrd 的结合标准误较个体标准误平均值大了约 70%。在表 5.3 中明显可以看出 rmbrd 系数具较大的变异，其估计值为 1.66 到 2.95。

图 5.3 中标示为“t for H0: Mean = 0”的字段就只是每一个系数对于其标准误的比。紧接着的字段给出了用于从 t 表中计算 p 值的自由度。这个数目与观察值的数目或变量的数目无关。虽然没必要知道自由度是如何计算的，但作者认为需要对该值做简短的解释。对于一个给定的系数，令 U 为回归模型中标准误平方的平均值。令 B 为回归间系数的方差。因缺失数据导致方差中相对的增加则被定义为

$$r = \frac{(1+M^{-1})B}{U}$$

其中 M 和之前相同，是用来产生估计值的完整数据组数目。自由度则被计算为

$$df=(M-1)(1+r^{-1})^2$$

因此，相对于回归内变化回归间变化越小的话，自由度越大。有时候，自由度将会远大于观察值的数目。但毋需担心，因为任何大于 150 左右的数目会造成一个实质上如同依标准正态分布的 t 表。然而，有些软件（包含 PROC MIANALYZE）可以产生一个不大于样本数目的、调整过的自由度（Barnard & Rubin，1999）。

最后一栏缺失信息比是因缺失数据导致每个系数有多少信息缺失的估计值。其范围比 lenroll 的低 21%，高于 rmbrd71%。对于有着 40%缺失数据的 rmbrd 而言，有高的缺失信息并不令人吃惊，但对于没有缺失数据的 private 及少于 1%缺失数据的 stufac 而言，缺失信息则高得令人吃惊。要理解这点，必须要先知道一些知识。首先，对于每一个给定系数，其缺失信息量不只取决于该特定变量的缺失数据，也取决于与其相关的其他变量之缺失数据百分比。其次，MIANALYZW 程序没法知道每一个变量有多少缺失数据。相反，缺失信息估计值全然立足于回归内和回归间的相对变化。如果在回归间有大的变化，则暗示有较多缺失信息。有时候缺失信息比以 γ 表示，用我们刚定义过的两个统计量 r 和 df 计算。即：

$$\hat{r}=\frac{r+2/(df+3)}{r+1}$$

同时我们要明确，在表中报告的缺失信息比可能只是一个受到大量抽样变化所影响的估计值。

如前所述，多重插补的麻烦之一是它不会产生一个确定的结果。每次执行多重插补，都会得到有些微差异的估计值和相关的统计量。为了理解这点，我们看一下图 5.4，其立足于由一个全新的数据扩增法运算所产生的 5 个数据组。虽然 lenroll 和 private 的缺失信息比之前的低很多，但大多数的结果与图 5.3 非常相似。

Multiple-Imputation Parameter Estimates

Variable	Mean	Std Error Mean	DF	t for H0: Mean=0	Pr>|t|	Fraction Missing Information
intercept	−32.474158	4.816341	124	−6.742496	<0.0001	0.192429
csat	0.066590	0.005187	20	12.838386	<0.0001	0.489341
lenroll	2.173214	0.546177	2157	3.978955	<0.0001	0.043949
private	13.125024	1.171488	1191	11.203719	<0.0001	0.059531
stufac	−0.190031	0.099027	51	−1.918988	0.0607	0.307569
rmbrd	2.357444	0.599341	12	3.993396	0.0020	0.623224

图 5.4 从 MIANALYZE 多重插补复制而得的输出结果

当缺失信息比较高时，多于平常建议的 3 到 5 个完整数据组可能是有必要的，从而得到稳定的估计值。那么到底需要多少数据组呢？一个有着无限数目数据组的多重插补为完全有效（如同 ML），但具有限数目数据组的 MI 不能达到全然有效。Rubin(1987)证明了一个根据 M 个数据组的估计值与一个根据无限数目数据组的估计值，两者之相对有效性为 $(1+\gamma/M)^{-1}$，其中 γ 为缺失信息比。这暗示有 5 个数据组和 50%的缺失信息，其估计程序的效率为 91%。有 10 个数据组的话，效率提升至 95%。同样，只有 5 个数据组得到的标准误比由无限数目的数据组所提供的标准误大 5%。10 个数据组产生的标准误比由无限数目的数据组所提供的标准误大 2.5%。其结果是，即使有着 50%的缺失信息，但 5 个数据组已表现得相当好了。将数据组数目扩大至其两倍，可以将过量的标准误减半，但此时的标准误已经很小就不需要再这样做了。

在结束回归实例前，让我们比较图 5.4 的 MI 结果和表 4.6 的 MI 结果。我们发现，系数估计值相当相似，标准误和 t 统计量也一样相当相似。毫无疑义，这两个分析会得出相同的结论。

第 6 章 | 多重插补:复杂化

MI 中的交互作用和非线性

虽然我们之前叙述过的方法对于估计有缺失数据的变量的主要效应有非常好的效果,但它们对于估计交互作用效果可能并不理想。例如,假设我们怀疑公立和私立大专院校,SAT 分数(CSAT)对于毕业率(GRADRAT)的效应有所不同。一个检验这个假设的方法(方法 1)是取之前插补的数据,创造一个等于 CSAT 和 PRIVATE 乘积的新变量,并将这个新变量与已在模型中的其他变量一起纳入回归方程中。

表 6.1(方法 1)最左边的一栏显示了这种做法的结果。变量 PRIVACAST 是 CSAT 和 PRIVATE 的乘积。p 值为 0.39,交互作用在统计上一点儿也不显著,所以我们可以总结为,在公立和私立机构间,CSAT 的效应没有改变。

表 6.1 有交互作用项的回归:三种方法

变 量	方法 1		方法 2		方法 3	
	系数	p 值	系数	p 值	系数	p 值
INTERCEPT	−39.142	0.000	−48.046	0.000	−50.2	0.000
CSAT	0.073	0.000	0.085	0.000	0.085	0.000
LENROLL	2.383	0.000	1.932	0.001	1.950	0.013
STUFAC	−0.175	0.208	−0.204	0.083	−0.152	0.091
PRIVATE	20.870	0.023	35.128	0.001	36.118	0.002
RMBRD	2.134	0.002	2.448	0.000	2.641	0.003
PRIVCSAT	−0.008	0.388	−0.024	0.022	−0.024	0.024

这个方法的问题是，虽然多元正态模型精于插补新产生变量间线性关系的值，但它并不模型化任何其他高阶动差。因此，除非使用特别的技巧，不然插补值不会显示具交互作用的证据。在这个例子中，交互作用两个变量中的其中一个(PRIVATE)是二分的，最自然的解法(方法 2)是对私立大专院校和公立大专院校分别做系列的数据扩增法。这允许 CSAT 和 GRADRAT 间的关系在这两个群体间有所差异，且允许插补值反映这个事实。一旦完成分别的插补，数据组会重新结合成一个单一数据组，乘积变量被创造，就可用该乘积变量来运行回归。表 6.1 中间部分的结果显示，PRIVATE 和 CSAT 间的交互作用在 0.02 程度上显著。我们可以更明确地发现，CSAT 对于毕业率的正效应，私立大专院校比公立大专院校小。

第三个方法(方法 3)对所有具 CSAT 和 PRIVATE 的观察值的个案，在插补前创造乘积变量，接着再如同任何其他有缺失数据的变量一样插补该乘积变量，最后使用插补数据估计包含该乘积变量的回归模型。

这个方法不如方法 2 那么吸引人，因为很明显乘积变量会有一个根本不是正态分布的，而在插补过程中却假设了正态性。然而，如同在表 6.1 右边部分所看到的，方法 3 的结果与方法 2 的结果非常接近，而且无疑比方法 1 更接近方法 2 的结果。

方法 3 的结果令人放心，因为当交互作用中的两个变量都被以一量化量尺测量时，方法 2 并不可行。因此，如果我们希望估计一个有 CSAT 和 RMBRD 的模型，我们需要对具这两个变量数据的 476 个个案创造一个乘积变量。对于其他剩余的 826 个个案，我们必须插补这个乘积变量作为数据扩增法过程的一部分。想估计一个牵涉具缺失数据的变量之非线性关系的模型时，就应该使用这个方法(或方法 2，当其为可行时)。例如，如果我们想要估计一个有 BMBRD 和 RMBRD 平方项两者的模型，则平方项必须被插补作为数据扩增法过程的一部分。这个需求会对在开始插补前就期待想要的函数形式的插补者造成一些负担。这也表示，我们必须从已被其他人使用的完全线性模型进行插补的数据中，谨慎地估计非线性模型。当然，如果一给定变量的缺失数据比例较小，我们选择用一个变量的原始形式插补它，随后再建构一个非线性的转换，可能会有

效果。毫无疑问，对于1302个个案，只有2个个案有缺失数据的STUFAC变量（学生与教职员比例），在将两个插补值平方后，将STUFAC平方项放入回归模型是可以接受的。

插补模型和分析模型之适合性

交互作用这个问题说明了一个在多重插补中更普遍的议题。比较理想的结果是，用来插补的模型与用于分析的模型一致，而且两者都正确地代表该数据。计算标准误基本公式（方程5.1）取决于它的适合性和正确性。

如果插补和分析的模型不同时会怎么样？这取决于差异的本质以及哪一个模型更加正确（Schafer，1997）。特别要关注的是那些在一个模型中为另一个模型特例的情况。例如，插补模型可以考虑交互作用，但分析模型可能不行，或者分析模型可以允许交互作用，但插补模型可能不行。这两个例子的其中任意一个，如果被较简单模型所施加的额外限制为正确的，那么我们已讨论过的在多重插补下的推论程序就是有效的。然而，如果额外的限制不正确，那么使用标准方法的推论可能就无效。

对于模型选择较不敏感的方法也被提议用来估计多重插补下的标准误（Wang & Robins，1998；Robins & Wang，2000）。很明显，当插补和分析的模型不兼容或者当两个模型都不正确时，这些方法给予有效的标准误估计值。然而，在任一阶段的不正确模型都仍有可能产生偏误的参数估计值，而且备选方法需要目前尚不能获取的专业软件。

插补中因变量所扮演的角色

因为GRADRAT是包含在数据扩增法过程中的变量之一，于是因变量暗中被用来插补自变量的缺失值。这样合理吗？它不会导致产生虚假的大的回归系数吗？答案是这不仅可行，而且为了要得到无偏误的回归系数估计值，这甚至是必要的。决定性的插补，使用因变量以插补自变量的缺失值，的确可能

产生虚假的大的回归系数，但将一个随机成分引入插补过程中，就可以抵消这个倾向并给予我们近似反映无偏误的估计值。事实上，将因变量排除于插补过程外，对于那些有缺失数据的变量而言，至少会倾向产生虚假的小的回归系数(Landerman, Land & Pieper, 1997)。在大专院校的例子中，如果 GRADRAT 不被用于插补，就会有大部分缺失数据的 CSAT 和 RMBRD，两者的系数分别会降低约 25%和 20%。同时，只有 5 个缺失值的 LENROLL 的系数会变大 65%。

当然，将 GRADRAT 纳入数据扩增法过程中也意味着 GRADRAT 的所有缺失值被插补了。有些学者就反对对因变量插补缺失值(Cohen & Cohen, 1985)。鉴于此项建议，我们需要在开始插补前，就去掉任何在因变量上有缺失数据的个案。这个建议有一个正当的基本理由，但它只在一些特别的例子中适用。如果因变量有缺失数据但所有的自变量没有缺失数据，则(不论线性或非线性的)回归模型之最大似然估计没有使用具缺失数据的个案的任何信息。因为 ML 是最合适的，在多重插补下插补缺失个案并不会获得任何好处。事实上，虽然诸如此类的插补不会导致任何偏误，但其标准误会比较大。然而，当自变量上也有缺失数据时，情况就不同了。因变量上有缺失数据的个案，确实有一些可以用于回归系数估计的信息，虽然信息量可能不是很大。结果是，在因变量和自变量两者皆有缺失数据的典型例子中，因变量上有缺失数据的个案不应该被删除。

在插补过程中使用额外的变量

如之前所提到的，用于数据扩增法的变量组确实应该包含那些将在计划分析中使用到的所有变量。在大专院校的例子中，我们也纳入一个额外变量 ACT (平均 ACT 分数)，因为其与有着大量缺失数据的变量 CSAT 高度相关。目的是要改善 CSAT 的插补以得到它更可靠的回归系数的估计值。如果我们包含其他与 CSAT 相关的变量，我们会做得更好。

一个相对较简单的例子就能阐明额外预测变量的好处。假设我们想要估计 1302 所大专院校的 CSAT 平均分数，我们知道，有 523 个个案有 CSAT 缺失

数据。如果我们使用有 CSAT 值的其他 779 个个案计算平均数，我们会得到表 6.2 第一行的结果。第二行则显示用多重插补及变量 ACT 的估计平均数(及标准误)。平均数降低了 9 个点，而标准误降低了 13%。虽然 ACT 和 CSAT 的相关性约为 0.90，但它作为一个预测变量的有用性，却被 523 个缺失 CSAT 的个案中只有 226 个个案有 ACT 的观察值这个事实所影响了。如果我们加入一个额外的变量 PCT25(班上前 25%的学生的比例)，我们的标准误会进一步减少。PCT25 与 CSAT 的相关性约为 0.80，且对于额外的 240 个 CSAT 和 ACT 两者皆缺失数据的个案，PCT25 更容易获取。

表 6.2　不同变量用于插补中所得的 CSAT 之平均数(及标准误)

用于插补中的变量	平均数	标准误	缺失信息%
没有	967.98	4.43	40.1*
ACT	956.87	3.84	26.5
ACT, PCT25	959.48	3.60	13.3
ACT, PCT25, GRADRAT	958.04	3.58	11.3

注：* 缺失数据的实际百分比。

表 6.2 最后一行加入了与 CSAT 相关性约为 0.60 的 GRADRAT，但只能利用 17 个未被 PCT25 或 ACT 所涵盖的个案。丝毫不出人意料，标准误的减少相当小。当作者试着引入图 5.4 中回归模型的所有其他变量时，标准误事实上变得更大。这可能是由于其他变量与 CSAT 的相关性低很多，然而因为需要估计他们的回归系数以预测 CAST，所以引进额外的变量。如同其他预测问题一样，当粗劣的预测量被加入模型时，插补可能会变得更糟。

多重插补的其他参数方法

如我们已经看到的，在广泛的数据类型和缺失数据形态下，多变量正态模型下的多重插补是相当简单的。作为处理缺失数据的一个惯例方法，这可能是当前最好的方法。然而，有数个备选方法在某些情况下可能更为可取。

多元正态模型最明显的缺点之一是，它只设计用来插补量化变量的缺失

值。如我们已经看到的，类别变量通过使用一些临时的方法来修正。然而，有时候你可能想做得更好。对于在插补过程中所有变量都是类别变量的情况下，有一个更好的模型，即无限制的多项模型（其在列联表中每一个单元格都有一个参数）或者一个允许在多项参数上有所限制的对数线性模型。在第4章中，我们讨论过这些模型的ML估计。Schafer(1997)也证明了这些模型如何能被使用以作为数据扩增法的基础，从而产生多重插补，而且他也开发了一个叫做CAT的免费软件程序以执行这个任务（参见 http://www.stat.psu.edu/~jls/）。

当数据包含类别的和量化变量的组合时，另一个 Schafer 程序（MIX）使用数据扩增法以产生插补。这个方法假设类别变量有一多项的分布，并在变量上可能有对数线性的限制。在由类别变量所创造出的列联表中的每一个单元格里，量化变量被假设有一个多变量正态分布。这些变量的平均数可以在各单元格变化，但协方差矩阵则被假设为固定的。

在作者写这本书时，CAT 和 MIX 还只能从 S-PLUS 统计软件包的程序库中获取，但它们之后可能会推出独立的版本。在之前的两个例子（CAT 和 MIX）中，基准模型包含的参数远远超过一般多元模型所包含的参数数目。因此，这些方法的有效运用很明显需要更多来自插补执行者的知识及创造，以及更大的样本数以达到稳定的估计值。

如果数据对于一个单一类别变量有缺失，logistic 回归模型下的多重插补就相当容易（Rubin, 1987）。假设在编码为5个类别的婚姻状态数据有缺失，且有数个可能的连续的和类别的预测变量。为了插补，我们使用具完整数据的个案，以一个预测量的函数，对婚姻状态估计一个多项的 logit 模型。这将会产生一组系数估计值 $\hat{\beta}$ 及一个协方差矩阵估计值 $\hat{V}(\hat{\beta})$。为了考虑参数估计值间的变异，我们从一个平均数为 $\hat{\beta}$ 且协方差矩阵为 $\hat{V}(\hat{\beta})$ 的正态分布中随机抽取参数估计值（Schafer 给出了如何有效实行这种操作的实用建议）。对于有缺失数据的每一个个案，抽取的系数值和被观察到的协变量值被代入多项的 logit 模型以产生落入5个婚姻状态类别的预测概率。根据这些预测概率，我们随机抽取1个婚姻状态类别，作为最终的插补值。[12]这整个过程被重复多次以

产生多个完整数据组。当然，一个二元变量只是这个方法的一个特殊例子。这个方法可被用于许多其他的参数模型，包含泊松回归及参数的失效时间回归。

无参数及部分参数方法

相对于我们刚考虑过的全然参数方法，在相对不严格假设的情况下，有人提议用另外几种方法来做多重插补。在这个部分，作者将考虑一些比较有代表性的方法，但需要注意的是，这些方法中几乎每一个都会存在许多变化情况。虽然当数据缺失为单调形态时，这些方法通常可以毫无困难地被普遍化到具多元变量的情况（在第 4 章已叙述过了），但这些方法在只有一个单一变量有缺失数据的情况下被运用最为自然。关于该内容可以参考 Rubin(1987)有关单调广义化的内容。当缺失数据不服从一个单调形态时，这些方法有时可以被使用，但在这样的设定下，它们明显缺少坚实的理论基础。

当在参数和非参数方法间选择时，通常要在偏误与抽样变异间有一个权衡。参数方法倾向于有较少的抽样变异，但如果参数模型对于所关注的现象并不是一个良好的近似反映时，它们可能会导致有偏误的估计值。非参数模型在许多情况下可能相对不容易出现偏误的情况，但估计值通常有更多的抽样变异。

热卡方法

最广为人知的非参数插补方法，是美国人口调查局经常使用的、同时也供公众使用的数据组插补值的“热卡”方法。

其基本理论是，我们想要插补某一特别变量 Y 的缺失值，该变量有可能是量化的或类别的。我们需要寻找一组与 Y 相关的、（没有缺失值的）类别变量 X。我们根据 X 变量制作一个列联表。如果在该列联表某个特定的单元格中个案有缺失 Y 值，我们取在同一个单元格中的一个或多个无缺失的个案，并用它们的 Y 值来插补这个缺失的 Y 值。

很明显，该方法可能会比较复杂，而且最重要的问题是如何选择"捐赠者"值以指派给有缺失值的个案？此外，捐赠者个案的选择应稍微被随机化从而避免偏误。这自然地导致多元回归，因为任何随机化方法可能不止一次地被运用以产生这个缺失值的不同的插补值。诀窍在于要做到随机化，使所有自然的变异都能被保留。为了达到这个目的，Rubin 提出了一个他创造的近似反映贝叶斯自举法(Rubin，1987；Rubin & Schenker，1991)。下面具体介绍如何使用这种方法。假设在列联表一特定的单元格中有 n_1 个个案于 Y 有完整数据且 n_0 个个案在 Y 有缺失数据。根据这些步骤：

(1) 从有完整数据的 n_1 个案组中，取一个 n_1 个案的随机样本(有取代)；

(2) 从这个样本中，取一个 n_0 个案的随机样本(有取代)；

(3) 指派 n_0 个的观察值给有缺失 Y 数据的 n_0 个个案；

(4) 对列联表中每一个单元格重复步骤 1 到 3。

当运用到列联表中所有的单元格时，这 4 个步骤会产生一个完整的数据组。对于多重插补，整个过程会被重复多次。在每一个数据组中都执行了所要的分析，用的同样是多变量正态插补的公式，同时所得结果加以汇总。

虽然我们似乎可以忽略步骤 1 而直接从有完整数据的 n_1 个案中选取 n_0 个捐赠者个案，但这无法估计标准误产生足够的变异。其他的变异是因为步骤 2 抽样用的是替代的方法。

预测均数匹配

热卡插补一个主要的吸引力在于，所有插补值都是实际被观察到的值。因此，没有"不可能出现的"或超出范围的值，且易于保存原分布的形状。其缺点为所有的预测变量必须是分类别的(或被当做如此)，因此对可能预测变量的数目进行了严格的限制。为了去除这个限制，Little(1988)提出了一个叫做预测均数匹配的部分参数方法。如同多变量正态参数方法一样，这个方法开始也基于有完整数据的个案，用 Y 对一组变量做回归，再由得到的回归模型产生有缺失和没缺失数据个案的预测值。接着，对于有缺失数据的每一个个案，我们找一组有完整数据的个案，其 Y 的预测值与有缺失数据个案的 Y

预测值是"接近"的。从这组个案，我们随机选择一个个案的 Y 值捐赠给该缺失个案。

对于单一变量 Y，可以直接将"接近"定义为预测值间的绝对差异。然而，接下来我们必须决定多少个接近的预测值应被包含进每一个缺失个案的捐赠库中？或相同地，构成可能的捐赠值组的接近的截略点应为何？如果选取了较小的捐赠库，在估计值中就会有较多的抽样变异性。另一方面，太大的捐赠库则容易导致可能的偏误，因为许多捐赠者可能与受援者不相像。为了处理这个模棱两可的情况，Schenker 和 Taylor(1996)提出了一个"适合的方法"，根据接近预测值的完整个案的"密度"，变化每一个缺失数据个案捐赠库的大小。他们发现，该方法比其他有固定的 3 个或 10 个捐赠库的方法稍微好些。然而，这三个方法间的差异太小以至于该方法几乎不值得我们去花额外的计算成本。

执行预测均数匹配时，因为回归系数只是真实系数的估计值，因此做些调整也是必要的。和参数的例子一样，这可以通过在计算每一个插补数据组的预测值前，先从它们的后验分布中随机抽取一组新的回归参数来完成。以下给出具体的操作步骤：

(1) 对于没有缺失 Y 数据的 n_1 个个案回归 Y 于 X(一协变量的向量)，产生回归系数 b(一个 $k\times 1$ 向量)及残差方差估计值 s^2；

(2) 从一个(假设不提供信息的先验的)残差方差的后验分布中进行随机抽样。可通过计算 $(n_1-k)s^2/\chi^2$ 来完成，其中 χ^2 代表从一个有 n_1-k 个自由度的卡方分布中随机抽取而得。令 $s_{[1]}^2$ 为第一个这样的随机抽取；

(3) 从回归系数的后验分布中随机抽取系数值。可通过从一个有着平均数为 b 及协方差矩阵 $s_{[1]}^2(X'X)^{-1}$ 的多变量正态分布抽取所完成，其中 X 为一个 X 值的 n_1-k 矩阵。此外，$b_{[1]}$ 为第一个这样的随机抽取。

有关如何执行的实用建议参见 Schafer(1997)。

对于每一组新的回归参数，所有个案产生预测值。接着对于每一个 Y 有缺失数据的个案，我们根据预测值建构一个捐赠者库，并随机从该捐赠库中选择

一个 Y 的观察值。虽然计算可能会变得相当复杂，但这个预测均数匹配法可以被广义化至超过一个 Y 变量有缺失数据的情况(Little, 1998)。

经验残差抽样法

在数据扩增法中，残差值从一个标准正态分布抽样而来，并且残差值接着被加到预测回归值中，以得到最终的插补值。我们可以通过在线性回归所产生的实际残差组中做随机抽取，来修正这个方法，从而使它较少依靠参数假设。这可以产生插补值，其分布更像该被观察到的变量的分布(Rubin, 1987)，虽然这个方法仍有可能得到落于允许区间外的插补值。

如同用其他方法进行多重插补一样，存在与正确执行这个方法涉及许多相关的重要的细节问题。和之前一样，令 Y 为有缺失数据的变量，用 n_1 个有观察数据的个案来插补 n_0 个个案。令 X 为没有缺失数据的 n_1 个个案之 $k\times 1$ 变量向量(包含一个常数项)。我们由执行前述的 3 个步骤开始，以获得 Y 对于 X 的线性回归，并且从参数的后验分布中产生随机抽取值。接着我们增加下列步骤：

(4) 根据步骤 1 中的回归估计值，对有缺失数据的个案计算标准化的残差：

$$e_i = (y_i - bx_i)/\sqrt{s^2(1-k/n_1)}$$

(5) 从步骤 4 计算得到的 n_1 个残差中抽取一个有 n_0 个值的简单随机样本(有替换)。

(6) 对于有缺失数据的 n_0 个个案，计算 Y 的插补值，如：

$$y_i = b_{[1]}x_i + s_{[1]}e_i$$

其中 e_i 代表步骤 5 中抽取的残差，而 $b_{[1]}$ 和 $s_{[1]}$ 是从参数的后验分布的第一次随机抽取得到的值。

这 6 个步骤产生一个完整的数据组。为了得到额外的数据组，简单地重复步骤 2 到 6 即可(除了不应该被重复的步骤 4 以外)。

Rubin(1987)曾经解释过，这个方法可以很容易被延伸至数个变量上有着单调缺失形态的数据组。当其为缺失时，每一个变量使用所有被观察到的变量

作为预测量来进行插补。这个经验残差方法也可以被修正以考虑插补值中的异方差性(Schenker & Taylor, 1996)。对于每一个插补的个案,残差库被限制为那些 Y 预测值与缺失数据个案的 Y 预测值接近的被观察到的个案。

实例

让我们对大专院校数据的一个次集合试试看部分参数方法。对 1272 所大专院校,TUITION 是全部都被观察到的(为了简便,我们应该去掉在该变量上有缺失数据的 30 个个案)。在这 1272 所大专院校中,只有 796 个报告 BOARD,即各大专院校的年度平均食宿费用。使用 TUITION 作为一个预测量,我们的目的是要对其他 476 所大专院校插补 BOARD 的缺失值,并且对所有 1272 所大专院校估计 BOARD 的平均数。

首先,我们运用先前使用过的方法。对于有完整数据成列删除的 796 所大专院校,BOARD 的平均值为 2060 美元,其标准误为 23.4。对 TUITION 和 BOARD 运用 EM 运算法,我们得到 BOARD 的平均数为 2032(但没有标准误)。BOARD 和 TUITION 间的 EM 相关性估计值为 0.555。使用数据扩增法、多变量正态模型下的多重插补产生的 BOARD 的平均数为 2040,估计标准误为 21.2。

因为 BOARD 高度右偏,故有理由怀疑多变量正态模型可能不是很恰当。相当多由数据扩增法所插补的值比最低的被观察值 531 小,且有一个插补值是负的。或许我们可以通过在经验残差上抽样从而做得更好。对于有 TUITION 和 BOARD 两者数据的 796 个个案,回归 BOARD 于 TUTION 的普通最小二乘(OLS)回归方程为:

$$\text{BOARD} = 1497.4 + 67.65 \times \text{TUITION}/1000$$

其均方根差估计为 542.6。这个回归方程计算得到的标准化残差由 796 个个案计算所得。

估计得到的回归参数被用于从参数的后验分布中做 5 次随机抽取,如同在步骤 2 和 3 中(假设一个不提供信息的先验分布)。被抽取的值为:

截　距	斜　率	均方根差
1536.40	66.6509	531.900
1503.65	71.5916	552.708
1501.61	66.9756	554.800
1486.84	66.9850	548.400
1504.23	61.2308	534.895

为了创造第一个完整数据组，476 个残差值从 796 个个案中被随机有替代的抽取而产生。这些标准化残差被任意地指派给 BOARD 有缺失数据的 476 个个案。令 E 为某一给定个案的指派残差，BOARD 的插补值产生如下

$$\text{BOARD} = 1536.40 + 66.6509 \times \text{TUITION}/1000 + 531.90 \times E$$

对 4 个剩余的数据组重复这个过程，并于每一个步骤中，在残差上有新的抽样，回归参数有新的值。

一旦产生 5 个数据组，就对每一个数据组计算其平均数和标准误，并使用标准误的方程 5.1 来计算结果。BOARD 的最终平均数估计值为 2035，估计标准误为 20.4，相当接近基于正态分布的多重插补。

现在，让我们尝试预测均数匹配。根据从 BOARD 对于 TUTION 的 OLS 回归的系数，作者从参数的后验分布中进行 5 次新的随机抽取：

截　距	斜　率	均方根差
1465.89	67.8732	557.531
1548.98	64.5723	539.952
1428.82	67.3901	512.381
1469.34	67.3750	550.945
1517.92	66.1926	534.804

对于第一组参数值，作者对所有观察到的及缺失的个案推测 BOARD 的预测值。对于每一个 BOARD 有缺失数据的个案，作者找到了预测值最接近缺失个案预测值的 5 个个案。同时随机选择这 5 个个案中的其中一个，并指派它被观察到的 BOARD 值作为插补值给缺失的个案。对 5 组参数值的每一组都重复

这个过程，以产生 5 个完整的数据组(数据组数目与被观察到的个案数目匹配到每一个缺失个案的数目相同只是巧合罢了)。接着对每一个数据组计算平均数和标准误并用通常的方法结合结果。BOARD 的结合平均数为 2028，其估计标准误为 23.0。

所有 4 个插补方法都产生相似的平均数估计值，且所有都显著低于根据成列删除而得的平均数。Schenker 和 Taylor(1996)建议，虽然参数和部分参数插补方法倾向产生相似的平均数结构(包含回归系数)的估计值，但它们可能对于插补值的边际分布产生更加不同的结果。他们的模拟研究显示，对于主要关注点是边际分布的应用，部分参数模型有一个明显的优点。当被用来产生估计值的回归以许多方式被错误设定时，更是如此。

连续的广义回归模型

数据扩增法的吸引力之一是，不像其他刚讨论过的参数或半参数方法，它可以很容易地处理有缺失数据的、有大量变量的数据组。遗憾的是，这个方法需要对所有变量指明一个多变量分布，而当变量有着多种类型的时候(例如，连续的、二元的及计数数据)，这并不是件容易的事。有另一种方法被提议用来处理有着数种不同变量类型的大的复杂数据组之缺失数据。不拟合一单一综合模型(例，多变量正态)，而是对每一个有缺失数据变量分别指明一个回归模型。这个方法涉及在数个回归模型中循环，在每一个步骤插补缺失值。

虽然这个方法非常吸引人，但它却不像我们已讨论过的其他方法那样具有很强的理论上的有效性。在本书写作之时，详细的论述只有 Brand(1999)、Van Buuren 和 Oudshoorn 以及 Raghunathan，Lepkowski，Van Hoewyk 和 Solenberger(1999)等人未发表的报告。在 Raghunathan 等人关于此方法的论述中，可利用的模型包含正态线性模型、二元 logistic 回归、多项 logit 及 poisson 回归。回归模型以一个特定顺序来估计，从有着最少缺失值的因变量进行到有最多缺失数据的因变量。我们用 Y_1 至 Y_k 来表示这些变量，并令 X 表示没有缺失数据的该组变量。

第一个“回合”的估计进行如下。回归 Y_1 于 X，并用相似于“对于多重插补的其他参数的方法”部分已叙述过的、对多项的 logit 模型产生插补值的方法以产生插补值。可以对插补值设定边界和限制。接着，回归 Y_2 于 X 和 Y_1，包含 Y_1 的插补值，并且产生 Y_2 的插补值。接着回归 Y_3 于 X、Y_1 和 Y_2（含两个 Y 的插补值），继续直到所有回归都被估计过了。第二个和后续的回合重复这个过程，但现在每一个变量要对所有使用从前面步骤产生的插补值的其他变量做回归。这个过程继续至某一个预先指定的回合数目或直到出现稳定的插补值。一个可以完成这些任务的 SAS 的宏命令可参见 http://www.isr.umich.edu/src/smp/ive。

Van Buuren 和 Oudshoorm 把他们的这个方法命名为基于链式方程的多重插补，而且他们开发了 S-PLUS 程序以执行它（参见 http://www.multiple-impuation.com/）。他们的方法和 Raghunathan 等人的方法的主要差异在于不包含泊松分布，但对于插补值的随机抽取方法却允许有更多的选择（包括参数和部分参数）。

线性假设检验和最大似然比检验

我们所使用的多重插补的统计推论方法一直都非常简单。对于一个给定参数，该估计值的标准误通过方程 5.1 来计算。这个标准误接着被插入立足于正态近似反映的传统公式中，以对某些关注的假设产生一个置信区间或一个 t 统计量。有时候，这样并不足够。通常我们想要对参数组检验假设，例如，两个参数相等或数个参数全部都等于 0。当我们对一组虚拟变量估计数个系数时，这些种类的假设显得特别重要。此外，有必要通过比较一个模型与另一个较简单的模型来计算似然比统计量。当执行多重插补时，完成这些任务并不是那么容易。Schafer(1997)叙述过 3 种不同的方法，但没有一种是完全令人满意的。作者在此简短地叙述这 3 种方法，同时也会在下一个部分给出一个实例。

使用结合协方差矩阵的 Wald 检验

当没有缺失数据时，一个关于多元参数推论的普遍方法为，根据参数估计

值及其估计的协方差矩阵来计算 Wald 卡方统计量。这里有简单的回顾，但遗憾的是，需要矩阵代数学。假设我们想要估计一个 $p\times 1$ 参数向量 β。我们有估计值 $\hat{\beta}$ 和估计的协方差矩阵 C，我们想要检验一个表达为 $L\beta = c$ 的线性假设，其中 L 为一个 $r\times p$ 的常数项矩阵，且 c 为一个 $r\times 1$ 常数项向量。例如，如果我们想要检验 β 的前两个元素彼此相等这个假设，我们需要 $L = [1\ -1\ 0\ 0\ 0\ \cdots\ 0]$ 及 $c = 0$。Wald 检验的计算如下：

$$W = (L\hat{\beta} - c)'[LCL'](L\hat{\beta} - c), \qquad [6.1]$$

其在零假设下，有一个自由度为 r 的近似反映卡方分布。[13]

现在我们广义化这个方法到多重插补的情形中。不用 $\hat{\beta}$，我们可以使用 $\bar{\beta}$，即横跨数个完整数据组的估计值之平均数，也就是，

$$\bar{\beta} = \frac{1}{M-1}\sum_k \hat{\beta}_k$$

其次我们需要一个结合样本内变异与样本间变异的协变量矩阵的估计值。令 C_k 为在数据组 k 中的参数估计的协方差矩阵，且令 $\bar{C}$ 为那些横跨 M 个数据组矩阵的平均。样本间变异被定义为

$$B = \frac{1}{M-1}\sum_k (\hat{\beta}_k - \bar{\beta})(\hat{\beta}_k - \bar{\beta})'$$

协方差矩阵的结合估计值则为

$$\tilde{C} = \bar{C} + (1 + 1/M)B$$

这只是方程 5.1 的没有平方根的一个多变量广义化。我们用包含 $\bar{\beta}$ 和 $\tilde{C}$ 的方程 6.1，来替代 $\hat{\beta}$ 和 C，以得到我们的检验统计量。

遗憾的是，在典型的 M 小于或等于 5 的例子中，这起不到好的作用。在诸如此类的例子中，B 是一个对于协方差矩阵相当不稳定的估计量，且造成的 W 的分布不是卡方。Schafer(1997)给予一个对协方差矩阵更加稳定的估计量，但这需要不合理的假设，即假设对于 $\hat{\beta}$ 的所有元素，缺失信息比都相同。然而，有些模拟研究显示，即使当假设被违反时，这个备选方法仍行得通。这个方法已

被纳入 SAS 程序 MIANALYZE 中。

似然比检验

如果感兴趣的模型通过最大似然来估计且没有缺失数据，通常通过计算似然比卡方检验来执行多参数检验。这个程序相当简单。令 l_0 为强加假设时模型的对数似然，并令 l_1 为放松假设时模型的对数似然。似然比统计量即为 $L = 2(l_1 - l_0)$。

与之前一样，我们的目的是要将这个广义化到多重插补中。第一个步骤是要对 M 个完整数据组中的每一个执行想要的似然比检验。令 $\bar{L}$ 为横跨 M 个数据组的似然比卡方的平均数，这是较容易的部分。接下来是比较困难的部分。为了得到那些卡方，必须要在每一个数据组中估计两个模型，即有强加假设的模型和放松假设的模型。令 $\bar{\beta}_0$ 为强加假设时，M 个参数估计值的平均数，且令 $\bar{\beta}_1$ 为放松假设时，M 个参数估计值的平均数。在每一个数据组中，我们接着计算一个参数值被限制为 $\bar{\beta}_0$ 的模型的对数似然，且再次计算一个参数值被设为 $\bar{\beta}_1$ 的模型的对数似然(这明显需要能够计算和报告使用者指定参数值的对数似然比的软件)。根据这两个对数似然，在每一个数据组中计算一个似然比卡方。令 $\tilde{L}$ 为这些横跨 M 个样本的卡方统计量的平均数。

最终的检验统计量则为 $\tilde{L}/\left(r+\left(\frac{M+1}{M-1}\right)(\bar{L}-\tilde{L})\right)$，其中 r 为假设所强加的限制数目。在零假设下，这个统计量有着近似反映分子自由度为 r 的 F 分布。分母自由度的计算有点困难。令 $t = r(M-1)$ 且令

$$q = \left(\frac{M+1}{M-1}\right)\left(\frac{\bar{L}-\tilde{L}}{r}\right)$$

如果 $t > 4$，则 d. d. f. $= 4+(t-4)[1+(1-2/t)/q]^2$。如果 $t \leqslant 4$，则 d. d. f. $= t(1+1/r)(1+1/q)^2/2$。

结合卡方统计量

Wald 检验和似然比检验，两者都缺乏前面使用的单一参数方法的简单性。

而且它们需要有专业选项及输出的分析软件，这是我们通常都尽量避免的。

作者现在讨论一个容易从标准输出所计算的第三个方法，但该方法可能不像另外两种方法那样精准(Li, Meng, Raghunathan & Rubin, 1991)。该方法所需要的仅是计算在 M 个完整数据组中，每一个数据组的传统的卡方统计量(Wald 或似然比)，以及相关的自由度。

令 d_k^2 为数据组 k 中有着自由度为 r 的卡方统计量，令$\bar{d}^2$为 M 个数据组的这些统计量的平均数，且令 s_d^2 为 M 个数据组的卡方统计量的平方根的样本方差，也就是

$$s_d^2=\frac{1}{M-1}\sum_k(d_k-\bar{d})^2$$

被提议的检验统计量为

$$D=\frac{\bar{d}^2/r-(1-1/M-1)s_d^2}{1+(1+1/M-1)s_d^2}$$

在零假设下，这个统计量有着近似反映分子自由度为 r 的 F 分布。分母自由度近似反映

$$\left(\frac{M-1}{r^{3/M}}\right)\left(1+\frac{M}{(M+1/M)s_d^2}\right)^2$$

作者已经写了一个 SAS 的宏命令(COMBCHI)以执行这些运算和计算一个 p 值。该命令可以参见作者的网站(http://www.ssc.upenn.edu/~Allison)。你只需要键入数个卡方值以及自由度，宏命令就会报告一个 p 值。

MI 实例 2

让我们考虑另一个详细的经验实例来说明本章所讨论过的某些技巧。数据组来自 1994 年综合社会调查，共有 2992 名受访者(Davis & Smith, 1997)。我们的因变量为 SPANKING，一个对于“有时候管教一个小孩责打是必要的。你强烈赞成、赞成、反对或强烈反对”这个问题的回答。如同问题本身所提示

的，有 4 个可能的依序的答案，被编码为 1 至 4 的整数。这个问题被设计作为只对随机的三分之二的样本施行的模块的一部分。因此，有 1015 个完全随机缺失的个案。此外，另有 27 个缺失个案，其回答被编码为“不知道”或“没有回答”。

我们的目的是要估计一个依序的 logistic（累进的 logit）模型（McCullagh，1980），其中通过下列变量预测 SPANKING：

AGE 受访者的年龄以岁计算，从 18 到 89。缺失 6 个个案。

EDU 受教育年数的数目。缺失 7 个个案。

INCOME 家庭收入，以 21 个区间类别的终点编码，以千元计算。缺失 356 个个案。

FEMALE 1 = 女性；0 = 男性。

BLACK 1 = 黑人；0 = 白人，其他。

MARITAL 婚姻状态的 5 个类别。缺失 1 个个案。

REGION 9 个区域类别。

NOCHILD 1 = 没有小孩；其他为 0。缺失 9 个个案。

一个额外的变量 NODOUBT 需要更进一步的解释。受访者被问到他们对上帝的信仰，有 6 个回答类别，从“我不信上帝”到“我知道上帝真的存在而且我对于这点没有疑惑”。有 62%的受访者的典型回答是“我知道上帝真的存在而且我对于这点没有疑惑”。然而，如同责打那个问题一样，这个问题只是问卷模块的一部分，只随机询问 1386 个受访者中的一部分。因此，根据设计有 1606 个缺失个案。其余 60 个缺失是因为他们说“不知道”或“没有回答”。如同这里所使用的，如果受访者“没有疑惑”变量被编码为 1，反之则会被编码为 0。

大多数的缺失数据在 3 个变量以上：SPANKING、NODOUBT 和 INCOME。样本中有 5 个主要的缺失形态，占了 96%的受访者：

771 个个案　在任何变量上都没有缺失

927 个个案　只有 NODOUBT 缺失

421 个个案　只有 SPANKING 缺失

89 个个案　只有 SPANKING 缺失

509 个个案　SPANKING 和 NODOUBT 缺失

421 个个案　NODOUBT 和 INCOME 缺失

如往常一样，数据分析最简单的方法是成列删除，其只使用 26%的原始样本。为了指明模型，作者对婚姻状态类别创造虚拟变量：NEVER(从未结婚)、DIVSEP(离婚或分居)及 WIDOW(鳏寡)，以已婚者为参照组。对区域也生成 3 个虚拟变量(以西方作为省略的类别)。[14](由 SAS 中 PROC LOGISTIC 所产生的)结果显示于表 6.3 的第一栏中。黑人、较老的受访者及对于上帝“没有疑惑”的受访者有可能赞同责打小孩有时是必要的。女性及受更多教育的受访者则可能比较反对。也有大的区域性差异：来自南方的受访者对于责打更加赞同，而来自东北的则更加反对。另一方面，没有证据显示收入、婚姻状态或有小孩等有任何效应。

表 6.3　预测 SPANKING 累进的 logit 模型的系数估计值(及标准误)

变　量	成列删除	正态数据扩增法	连续回归	连续回归(缺失数目于 SPANKING 上)
FEMALE	−0.355(0.141)*	−0.481(0.089)***	−0.449(0.098)***	−0.489(0.094)***
BLACK	0.565(0.218)**	0.756(0.117)***	0.693(0.119)***	0.685(0.135)***
INCOME	−0.0036(0.0033)	−0.0052(0.0020)*	−0.0042(0.0027)	−0.0047(0.0022)*
EDUC	−0.0055(0.027)*	−0.061(0.016)***	−0.073(0.019)**	−0.068(0.016)***
NODOUBT	0.454(0.147)**	0.465(0.120)**	0.455(0.156)*	0.438(0.121)**
NOCHILD	−0.205(0.199)	−0.109(0.112)	−0.141(0.164)	−0.091(0.123)
AGE	0.010(0.005)*	0.0043(0.0032)	0.0031(0.0032)	0.0040(0.0031)
EAST	−0.712(0.219)**	−0.444(0.125)***	−0.519(0.156)**	−0.488(0.136)***
MIDWEST	−0.122(0.203)	−0.161(0.136)	−0.228(0.149)	−0.159(0.128)
SOUTH	0.404(0.191)**	0.323(0.156)*	0.262(0.129)*	0.357(0.121)**
NEVMAR	−0.046(0.238)	−0.075(0.148)	−0.036(0.173)	−0.071(0.151)
DIVSEP	−0.191(0.194)	−0.203(0.150)	−0.141(0.128)	−0.184(0.126)
WTDOW	0.148(0.298)	−0.244(0.150)	−0.116(0.177)	−0.215(0.174)

注：* $p<0.005$，** $p<0.01$，*** $p<0.001$。

因为依据设计有 84%观察值有缺失数据(且因此为完全随机的)，成列删除

应该产生近似反映无偏误的估计值。然而，损失几乎四分之三的样本是要付出重大代价的，而这个代价若使用多重插补则是可以避免的。为了执行 MI，作者首先使用第 5 章叙述过的多变量正态模型下的数据扩增法。执行之前，先删除婚姻状态有缺失值的个案，以避免必须插补一个多类别变量。对去掉缺失数据的 SPANKING 变量的 1042 个个案，有一个合理的论点可以被提出，因为因变量上有缺失的个案包含很少关于回归系数的信息。然而，包含它们并没有损失而且可能还会有一些好处，所以作者把他们全部保留了。在模型中所有 13 个变量被全部纳入插补过程，没有经过任何正态化的转换。

所有虚拟变量的插补值被四舍五入至 0 或 1。SPANKING 的插补值被四舍五入至 1 到 4 的整数。年龄和收入有一些正当范围外的插补值，而这些值被重新编码至上限或下限。累进的 logit 模型接着被用来对 5 个数据组通过标准公式进行估计，得到的估计值再合并在一起。

结果显示在表 6.3 的第二栏中。结果的基本模式是相同的，这一栏中 INCOME 的效应是显著的，但 AGE 变得不显著了。最引人注目的是，所有系数的标准误都比那些成列删除的标准误低很多，最典型的低了约 40%。甚至 NODOUBT 的标准误也小了 18%，这是非常惊人的，因为超过一半的个案在该变量有缺失。对于许多变量较小的标准误产生低很多的 p 值。

累进的 logit 模型强加一个被称为成比例发生比假设以限制数据。简单地说，这个词组指对于任何因变量，其二分化系数被假设为相同的。PROC LOGISTIC 对成比例发生比假设为正确的这个零假设，报告一个卡方统计量(分数检验)。但由于我们处理 5 个数据组，因而得到 5 个卡方：32.0、31.3、38.0、36.4 和 35.2，每一个自由度都为 26。使用前述宏命令 COMBCHI，这 5 个值被结合以产生一个 0.25 的 p 值，暗示模型强加的限制能很好地拟和数据。对于这 5 个数据组中的每一个，作者也计算了所有区域系数都为 0 的零假设下的 Wald 卡方。自由度为 3，这些 Wald 卡方值为 72.9、81.3、53.4、67.7 和 67.0。结合的 p 值为 0.00002。

在表 6.3 的第三栏中，我们看到运用 Raghuanthan 等人(1999)的多重插补方法的结果，这些结果依靠连续的广义模型。对于每一个有缺失数据的变量估

计一个回归模型，将该变量当做因变量，其他所有变量当做预测量。这些回归模型接着被用来产生5组随机插补值。对EDUC和INCOME而言，虽然上下限被嵌入插补过程中，但该模型为普通线性模型。Logistic模型则指明给NODOUBT和NOCHILD。一个多项logit模型被用于SPANKING。在插补过程中有20个回合，也就是说，对于5个完整数据组中的其中任意一个，直到得出最终结果前，有缺失数据的变量会被连续地插补20次。

同之前一样，累进的logit模型被用来估计5个完整数据组中的每一个，并使用公式结合结果。表6.3第三栏中的系数估计值与那些用多变量正态数据扩增法得到的系数估计值相当类似。标准误通常会比数据扩增法的标准误大，虽然不像成列删除的标准误那么高。

令人意外的是，连续回归的成比例发生比假设卡方统计量几乎是正态数据扩增法的成比例发生比假设的卡方统计量的两倍大。更明确的是，自由度为26，其值为54.9、59.9、66.7、85.4和59.0，每一个都有一个小于0.001的p值。然而，当用宏命令COMBCHI结合这些值时，得到的p值为0.45。为什么个别的p值和结合p值间会有这么大的差异呢？其答案在于，在卡方间的大的方差表示它们中的每一个可能都是严重高估的值。结合它们的公式已经考虑到这点了。

那么正态扩增法和连续回归间卡方的不同又如何解释呢？作者怀疑这源自插补SAPNKING的多项的logit模型没有对该变量强加任何顺序的事实。因此，插补值不可能与成比例发生比假设相符。当以一个SPANKING的线性模型重做连续插补时（将插补值四舍五入到整数），成比例发生比假设的卡方与那些在正态数据扩增法下获得的卡方更加一致。二者择其一，作者在先删除所有SPANKING缺失个案后重做连续插补。SPANKING仍然被指名为类别的，亦即当插补其他变量的值时，它被视为一个类别的预测量。而且，成比例发生比假设的卡方跟那些由正态数据扩增法产生的卡方相似。

表6.3的最后一栏显示删除SPANKING缺失个案后连续回归插补的结合结果。有趣的是，两者系数和它们的标准误普遍地更接近于数据扩增法。更进一步讲，当我们删除1042个SPANKING有缺失数据的个案时，并没有明显的

信息损失。

长期的及其他集群数据的 MI

到目前为止，我们已假设每一个观察值独立于其他观察值，如果数据为每个大总体中的简单随机样本，这即为一个合理的假设。然而，许多数据组可能在各观察值间有某种相关性。例如，假设我们有一个面版个体数据，对他们而言，连续五年每年都测量相同的变量。许多分析面版数据的计算机程序需要组织数据，将每一年的测量当做个别的观察值。为了将观察值连结在一起，必须有一个变量包含识别号码，而识别号码对于来自相同个体的所有观察值都是一样的。

因此，如果我们有 100 个连续五年被观察的个体，我们会有 500 个有效的观察值。显而易见，这些观察值不会是独立的。如果已讨论过的多重插补方法被直接运用于这 500 个观察值上，不会利用到任何长期的信息。因此，完整数据组可能会产生长期相关性的、严重的低估值，特别是如果有大量的缺失数据的话。

如果观察值自然落入自然发生的集群中，也会产生相似的问题。假设我们有一个 500 对已婚者的样本，且对夫妻二人都询问相同的问题。如果我们对配偶中任选其一插补缺失数据，那么使用配偶间答案的相关性是很重要的。相同的也适用于在同一教室内的学生或同一小区内的受访者。

一个对于这些数据类别的处理方法是，在一个嵌入观察值间相关性的模型下进行多重插补。Schafer(1997)对集群的数据提出了一个多变量线性混合效应模型，并且也开发了一个计算机程序(PAN)以使用数据扩增法执行插补(参见 http://www.stat.psu.edu/~jls/)。虽然之前承诺会有一个 Windows 版本，但目前仅以 S-PLUS 套装软件的一个程序库来运行。

有一个更简单的方法，当调查的次数相对较少时，对于跟踪数据可以处理得很好。基本的想法是要格式化数据，以致对于每一个个体只有一个记录，而且在不同时点对同一变量有不同的测量。多重插补接着以我们所考虑过的任何一种方法来执行。这就把在任何时点的变量，用来作为任何其他时点的变量

的预测量的情况考虑进去了。一旦数据被插补，数据组就可以被重新格式化，以致对于每一个个体都有数笔记录，其对每一个时点都有一笔记录。

MI 实例 3

这里有一个比较简单的使用刚讨论过的方法进行长期数据多重插补的例子。样本由 220 名白人女性所组成，年龄至少都在 60 岁以上，在大宾州地区进行髋关节骨折手术(Mossey, Knott & Craik, 1990)。在她们出院后，她们被访问 3 次：2 个月后、6 个月后及 12 个月后。下列 5 个变量在 3 次访问中每一次都会被测量。

CESD　一个忧郁程度的测量，范围从 0 至 60。

SRH　自我健康评估，以一个包含 4 个值的量表测量(1＝不良，4＝杰出的)。

WALK　如果病患在家不用帮助可以行走，编码为 1；反之为 0。

ADL　可以不用协助而完成的自理"日常生活活动"的数目。

PAIN　病患所经历的疼痛程度范围，从 0(没有)到 6(持续)。

我们的目的是要以 CESD 作为因变量及其他 4 个作为自变量来估计一个"固定效应"线性回归模型(Greene, 2000)。模型的形式为

$$y_{it} = \alpha_{it} + \beta_1 x_{it1} + \cdots + \beta_4 x_{it4} + \varepsilon_{it}$$

其中 y_{it} 为个人 i 在时点 t 的 CESD 值，而 ε_{it} 满足线性模型的通常假设。这个模型值得注意的是，对于样本中每一个个人，都有一个不同的截距 α_i，从而控制病人们所有稳定的特性。这个个人特定的截距也引入了每一个个人的多个回答间的相关性。

为了估计模型，还创造了一个有 660 个观察值的可用的数据组，对每一个时点每一个个人都有一笔观察值。要得到 OLS 回归估计值有两个相同的计算方法：(1)对于每一个个人(小于 1)包含一个虚拟变量；(2)执行回归于离差分数上。第二个方法在执行多元回归前，先在模型中用每一个变量减去(跨三个时

点的)个人特定的平均数。

遗憾的是,该研究有大量的流失,以及各个时点上额外的无回答。如果我们删除所有具任何缺失数据的个人时点,可用的数据组会从 660 个个案减少至 453 个个案。如果我们删除在任何时间任何变量有缺失数据的个人,则数据组会减少为 101 个个人(或 303 个个人时点)。

表 6.4 显示使用 4 个方法处理缺失数据的固定效应回归结果。[15]前两栏给出了从两种成列删除版本得来的系数和标准误:(1)删除有任何缺失数据的个人;(2)删除有任何缺失数据的个人时点。有一个明显的证据显示,忧郁程度被自我健康评估所影响,对于行走能力的影响则不足为凭。忧郁程度在第 1 和第 2 次调查中明显地比第 3 次(当大部分的病患已完全痊愈)高出许多。很少或没有证据显示 ADL 和 PAIN 有效应。

表 6.4 预测 CESD 的固定效应模型之系数估计值(及标准误)

	成列删除,以人	成列删除,以个人时点	数据扩增法,以个人时点	数据扩增法,以个人
SRH	2.341(0.586)**	1.641(0.556)**	2.522(0.617)**	1.538(0.501)**
WALK	−1.552(0.771)*	−1.381(0.761)	−1.842(0.960)	−0.550(0.825)
ADL	−0.676(0.528)	−0.335(0.539)	−0.385(0.562)	−0.410(0.435)
PAIN	0.031(0.179)	0.215(0.168)	0.305(0.180)	0.170(0.164)
WAVE 1	8.004(0.650)**	8.787(0.613)**	6.900(0.729)**	9.112(0.615)**
WAVE 2	7.045(0.579)**	7.930(0.520)**	5.808(0.642)**	8.131(0.549)**
N(个人时点)	303	453	660	660

注:* $p<0.05$, ** $p<0.01$。

最后两栏给出了根据全部样本,含已在多变量正态模型下以数据扩增法插补的缺失数据的分析结果。[16]第三栏的结果,对被视为是独立观察值的 660 个人时点进行插补。因此,缺失数据仅通过同一时点的信息来插补。为了进行最后一栏的插补,数据被重新组织为 220 个个人,在每一个时点有区分的变量名称。这样一来,每一个有缺失数据的变量根据所有三个时点的信息以进行插补。原则上,这应该产生比较好的插补,特别是因为一个缺失值可以被不同时点的同一变量的测量值所预测。

事实上，最后一栏所有的估计标准误都比倒数第二栏的小一些。它们也比两种成列删除方法中的任一种更小一些。另一方面，根据个人次数时点的数据扩增法之标准误则比两种成列删除方法稍微大些。无论如何，在这个应用中，多重插补并不具有压倒性的优势。从质量上讲，不管使用何种插补方法，结论几乎都是相同的。

第 7 章 | 不可忽略的缺失数据

前面几章着重讲述了在缺失数据机制可以忽略的情况下能够使用的方法。可忽略性意味着我们不需要对数据发生缺失的过程进行模型化。可忽略性重点要求数据是随机缺失的——一特定变量缺失数据的概率不取决于该变量的值(扣除分析中其他变量的作用之后)。

处理可忽略的缺失数据的基本策略可以简单地总结如下:调整所有缺失和非缺失数据间的可观察到的差异,并且假设所有剩余的差异为无系统性的。当然,这是一个熟悉的策略。标准回归模型就被设计来做这个——调整可观察到的差异并假设所有未观察到的差异为无系统性的。

遗憾的是,通常我们有足够的理由怀疑数据不是随机缺失的。例如,常识告诉我们,曾被逮捕过的人比不曾被逮捕过的人更不可能报告他们的逮捕状态,有高收入的人可能不会报告他们的收入。在临床药物试验中,变得更差的人比那些变得更好的人,更有可能退出临床药物试验。

在这些情况下应该怎么办呢?有处理不可忽略的缺失数据的模型且想要运用它是很自然的。然而,意料之中的是,很少有可用的软件可以估计不可忽略的模型(除了一个重要的例外——Heckman 的选择性误差模型)。其基本问题在于,对数据给定一个模型,只有唯一一个可忽略缺失数据机制,但有无限多不同的不可忽略的缺失数据机制。所以很难编写计算机程序,以处理即使是这些可能性的一小部分。此外,根据选择的模型,答案可能会变化很大。所以选择正确的模型非常重要,而该选择需要对所调查的现象有非常精准及详尽的知识。更糟糕的是,没有经验方法可以从一个不可忽略的模型(或从一个可忽略的模型)中分别出另一个“不可忽略的模型”。

也许你不会极端到说:"不要做那个",但你可能会这么说:"如果你做那个,要极端小心。"此外,如果你没有很多统计方面的专业知识,那么请找一个具备统计专业知识的合作者。本章针对处理不可忽略的缺失数据的一些方法,为大家提供一个简短的导论及概要。

你所需要明确的第一件事情是,作者已经介绍过的对于可忽略的缺失数据的方法——最大似然和多重插补——可即刻适用于处理不可忽略的缺失数据。假设选择的模型正确,那么这两个方法有如他们在可忽略的设定下的相同的最适特性。第二点要记住的是,任何有关不可忽略的缺失数据的方法应该要伴随一个敏感性分析。因为根据假设的模型,结果可能变化很大,故试验一貌似有理范围的模型并看它们是否产生相同的结果是很重要的。

两种模型

不管你选择最大似然还是多重插补,处理不可忽略的缺失数据有两种截然不同的方法:选择模型和形态混合模型。对一个缺失数据的单一变量解释最为容易。令 Y 为关注的变量且令 R 为一个虚拟变量,如果 Y 被观察到,其值为 1,如果 Y 缺失,则其值为 0。令 $f(Y, R)$ 为这两个变量的联合概率密度函数。选择一个模型意味着对 $f(Y, R)$ 选择某些明确的指定。

可以使用两种不同的方法因子化联合 p. d. f(Little & Rubin, 1987)。在选择模型中我们使用

$$f(Y, R) = \Pr(R \mid Y) f(Y)$$

其中 $f(Y)$ 为 Y 的边际密度,且 $\Pr(R|Y)$ 为给定某 Y 值下的 R 的条件式概率。用文字表示,如同没有缺失数据一样我们首先对 Y 建立模型。例如,我们可以假设 $f(Y)$ 为有着平均数 μ 和方差 σ^2 的正态分布,且 $\Pr(R|Y)$ 为

$$\Pr(R = 1 \mid Y) = \begin{cases} p_1 & if\ Y > 0, \\ p_2 & if\ Y \leqslant 0. \end{cases}$$

这个模型被确认且可以被 ML 所估计。对应形态混合模型的备选的联合 p. d. f

因子化公式为，

$$f(Y, R) = f(Y \mid R)\Pr(R)$$

其中 $f(Y, R)$是条件为是否缺失的 Y 的密度。例如，我们可以假设 $\Pr(R)$就是某个固定的 θ，且是有着方差为 σ^2 和当 $R=1$ 时平均数为 μ_1、而当 $R=0$ 平均数为 μ_0 的正态分布。遗憾的是，这个模型不被确认，因此若对参数没有更进一步的限制则无法估计。

形态混合模型似乎是一个思考缺失数据机制的不自然的方法。而且非常典型，我们假设数据的值(在这个例子中为 Y)是预先设定的。那么，根据数据收集程序，Y 的值可能对于我们是否确实获得想要的信息有一些影响。这样的想法与选择模型一致。另一方面，形态混合模型似乎倒转了因果关系的方向，允许缺失性影响所关注变量的分布。当然，关于因果关系方向的条件式概率是不可知的，其结果是，形态混合模型有时候比理论上更吸引人的选择模型更容易处理，特别是对于多重插补。作者下面将会给出选择模型和形态混合模型的一些例子。

Heckman 的样本选择误差模型

Heckman(1976)的样本选择误差模型为缺失数据选择模型的经典例子。该模型设计用于因变量在一个线性回归模型中有一些个案有缺失、但另一些个案没有缺失的情况。一个经常使用的例子为预测女性工资的回归，而对于不在劳动力市场中的女性，其工资数据必然是缺失的。很自然地，我们会假定如果她们的工资很低的话，她们就不可能进入劳动力市场。因此，数据并不是随机缺失。

Heckman 以潜变量公式化其模型，但作者将用一个更直接的设定来处理。对于一个 n 个个案的样本 ($i=1, \cdots, n$)，令 Y 为一方差为σ^2 的正态分布变量，且其平均数为

$$E(Y_i) = \beta X_i, \qquad [7.1]$$

其中，X_i 为自变量（包含一个截距为 1）的栏向量 M，而 β 为一个系数的列向量。目的是要估计 β。如果所有的 Y_i 都被观察到了，我们可以通过普通最小二乘回归得到 β 的 ML 估计值。然而，有些 Y_i 是缺失的。Y_i 上缺失数据的概率被假设服从 probit 模型

$$\Pr(R_i = 0 \mid Y_i,\ X_i) = \Phi(\alpha_0 + \alpha_1 Y_i + \alpha_2 X_i), \qquad [7.2]$$

其中 $\Phi(\cdot)$是一个标准正态变量的累进分布。除非 $\alpha_1 = 0$，否则数据不是随机缺失的，因为缺失的概率取决于 Y。

这个模型被确认（即使没有 X_i 或当 X_i 不进入该 probit 方程时），且可以被最大似然所估计。对于一个有 Y 的观察值，其似然为

$$\Pr(R_i = 1 \mid y_i,\ x_i) f(y_i \mid x_i) = [1 - \Phi(\alpha_0 + \alpha_1 y_i + \alpha_2 x_i)]\varphi\left(\frac{y_i - \beta x_i}{\sigma}\right)\sigma^{-1}, \qquad [7.3]$$

其中 $\varphi(\cdot)$为一标准正态变量的密度函数。对于一个有 Y 缺失的观察值，其似然为

$$\int_{-\infty}^{+\infty} \Pr(R_i = 0 \mid y_i,\ x_i) f(y \mid x_i) dy = \Phi\left(\frac{\alpha_0 + (\alpha_1\beta + \alpha_2)x_i}{\sqrt{1 + \alpha_1^2\sigma^2}}\right) \qquad [7.4]$$

方程 7.4 服从普遍原则，即有缺失数据的观察值，其似然可以通过求所有缺失数据可能值的似然之积分而得到。使用标准数值的方法，整个样本的似然值很容易被最大化从而作出估计。

遗憾的是，这个方法所产生的估计值对于 Y 为正态分布这个假设过度敏感。如果 Y 事实上有一个偏的分布，在 Heckman 模型下得到的 ML 估计值会严重偏误，或许甚至比从一可忽略的缺失数据模型下所获得的估计值偏误更加严重（Little & Rubin，1987）。

Heckman 也提出一个两步骤估计量，其对正态性的偏离不敏感，本质上更加容易计算，因此比 ML 更加受欢迎。然而，该两步骤方法有其自身的限制。

简而言之，两步骤如下：

(1) 估计一个缺失数据指标 R 的 probit 回归于 X 变量上；

(2) 对于 Y 有数据呈现的个案，估计一个最小二乘线性回归 Y 于 X 另加一个转换来自 probit 回归中的预测值之变量。[17]

不像 ML 方法，如果没有 X 变量，则两步骤程序不可行。此外，如果 X 变量在 proibt 和线性回归中都相同，参数只能被勉强地确定。为了得到合理稳定的估计值，在 probit 回归中的 X 变量必须从线性模型中排除。当然，能够令人信服地证明诸如此类的排除限制是很罕见的。甚至当所有条件都符合时，两步骤估计量在真实情况下也有可能表现得比较拙劣(Stolzenberg & Relles, 1990, 1997)。

在这些样本选择方法对于假设的违反有明显敏感性的前提下，我们应该如何继续做一个敏感性分析呢？对于 ML 估计量，关键在于假设为因变量 Y 的正态性。所以一个自然的方法是拟和假设不同分布的不同模型。偏的分布，如 Weibull 或 gamma 可能会是最有用的，因为对 ML 来说，正态分布的对称性最为重要。虽然在方程 7.4 中，积分可能没有一个合适的形式且可能需要数值的积分法，但是对于其他备选分布，ML 估计应该是可行的。对于两步骤估计量，关键假设是从预测的线性回归中排除某些 X 变量。一个敏感性分析可能探索对两个方程选择不同组 X 变量的结果。

形态混合模型的 ML 估计

形态混合模型很难被识别。假设我们有 2 个变量 X 和 Y，有 4 个观察到的缺失形态：

(1) X 和 Y 两者都被观察到；

(2) X 被观察到，Y 缺失；

(3) Y 被观察到，X 缺失；

(4) X 和 Y 两者都缺失。

根据这些形态中哪一种被观察到，令 $R = 1, 2, 3$ 或 4。对于这些数据的形态混合模型有一个普遍形式

$$f(X, Y, R) = f(Y, X \mid R)\Pr(R)$$

为了使模型更加明确，我们可以假设 $\Pr(R)$ 由 p_1，p_2，p_3 和 p_4 该组值所确定。接着，我们可以假设 $f(Y, X|R)$ 为一个有着通常参数 μ_X，μ_Y，σ_X，σ_Y，σ_{XY} 的二变量正态分布。然而，我们确定这些参数中的每一个对于每一个值是不同的。问题是，当 X 被观察到而 Y 没有被观察到时，没有信息可以估计 Y 的平均数和标准差或 X 和 Y 的协方差。同样地，当 Y 被观察到而 X 没有被观察到时，没有信息可以估计 X 的平均数和标准差或 X 和 Y 的协方差。如果两个变量都缺失，我们就没有任何信息。

为了继续下去，我们必须强加一些限制在这 4 组参数上。令 $\theta^{(i)}$ 为形态 i 的参数组。一个简单但非常有限制的条件是，假设 $\theta^{(1)} = \theta^{(2)} = \theta^{(3)} = \theta^{(4)}$，其相当于 MCAR。既然那样，形态混合模型的 ML 估计等同于第 3 章讨论过的、有可忽略的缺失数据的、正态模型的 ML 估计。Little(1993，1994)提出了对其他不对应但可忽略的缺失数据产生确认的模型的限制种类。这里有一个例子。令 $\theta^{(i)}_{(Y|X)}$ 代表在 Y 为形态 i 下，Y 的条件式概率。Little 称之为完全个案缺失变量限制：

$$\theta^{(2)}_{Y|X} = \theta^{(1)}_{Y|X},$$
$$\theta^{(3)}_{X|Y} = \theta^{(1)}_{X|Y},$$
$$\theta^{(4)} = \theta^{(1)}$$

对于两个具一个缺失变量的形态，在给定该观察到的变量下，该缺失变量的条件式概率等同于完整个案形态相对应的分布。对于有两个变量缺失的形态，所有参数被假设与在完整个案形态中的那些参数相同。这个模型被确认并可以以非迭代的方法得到 ML 估计值。一旦所有这些参数都能获得，就可以很容易地结合它们以得到 X 和 Y 的边际分布的估计值。

形态混合模型的多重插补

形态混合模型的 ML 在目前仍是相当难以理解的。更加实际且有用得多

的是组合形态混合模型与多重插补(Rubin, 1987)。最简单的策略是,首先在一个可忽略的模型下产生插补,接着以比如说一个线性来转换修正插补值。之后就可以很容易通过在线性转换中用不同的常数重复该过程来获得一个敏感性分析。

这里有一个简单的例子。再一次假设我们有两个变量 X 和 Y,但只有两种缺失数据形态:(1)完整个案,(2)Y 缺失。我们假设在每一种形态内,X 和 Y 有一个二变量正态分布。我们也相信有缺失的个案倾向于有较高的 Y 值,所以我们假设对两种形态所有参数都是相同的,除了 $\mu_Y^{(2)} = c\mu_Y^{(1)}$,其中 c 为某个大于 1 的常数。多重插补接着产生在一个可忽略的缺失数据机制下的 Y 的插补值,然后将所有插补值乘以 c。当然,要保证正常运行,我们要选一个 c 值,且该选择可以是随心所欲的。敏感性分析由对于数个不同 c 值,重插补数据及重估计模型所组成。[18]

现在,让我们将这个变成一个实际的例子。对于大专院校数据,有 98 所大专院校在因变量 GRADRAT 上有缺失数据。假设那些没有报告毕业率的大专院校,相对于那些有报告毕业率的大专院校,有较低的毕业率,是貌似合理的。我们在第 5 章叙述过的、对多个插补数据的插补毕业率的平均数比没有缺失数据的大专院校的平均毕业率低了约 10 个百分点这个事实,支持这项假设。然而,这个差异完全来源于预测变量的差异,而且这并不构成显示数据不是随机缺失的证据。

表 7.1　毕业率回归于数个变量上的不同形态混合模型

变　量	100%	90%	80%	70%	60%
CSAT	0.067	0.069	0.071	0.072	0.071
LENROLL	2.039	2.062	2.077	2.398	2.641
PRIVATE	12.716	12.542	11.908	12.675	12.522
STUFAC	−0.217	−0.142	−0.116	−0.216	−0.113
RMBRD	2.383	2.264	2.738	2.513	2.464

然而,假设在缺失及没有缺失个案间,毕业率的差异甚至会更大。毫无疑问,修正插补毕业率以使它们等于在可忽略性假设下被插补的指明的百分比。

表 7.1 给出了插补值为原始值 100%、90%、80%、70%和 60%的结果。对每一个回归,会产生全新的插补。因此,横跨各栏的变化是由插补过程的随机性导致的。一般而言,系数是相当稳定的,暗示对可忽略性的偏离对结论不会有很大的影响。STUFAC 系数变化最大,但在所有例子中它在统计上根本不显著。

第 8 章 | 总结与结论

在处理缺失数据的传统方法中，成列删除的问题最少。虽然成列删除可能会丢弃一大部分的数据，但除非数据不是完全随机缺失的，不然没有理由期望有误差。此外，其标准误也应该是真实标准误的适当的估计值。更进一步讲，如果你估计一个线性回归模型，对于一个自变量有缺失数据且缺失概率取决于该变量的值的情况下，成列删除是相当稳定的。如果你估计一个 logistic 回归模型，成列删除可以容忍因变量的非随机缺失或自变量的非随机缺失(但两者不能同时出现)。

相比之下，所有其他处理缺失数据的传统方法会在标准误估计值中引入误差，而且当数据为完全随机缺失时许多传统方法(如虚拟变量调整)仍会产生有偏误的参数估计值。所以成列删除是一个较为安全的方法。

如果在成列删除中必须被丢弃的数据量是无法容忍的，则有两种可备选的方法——最大似然法和多重插补。这两种方法假设数据随机缺失，这是一个较完全随机缺失更令人欣赏的较弱的假设。在颇为普遍的条件下，这些方法产生近似反映无偏误的、有效的估计值。它们也产生良好的标准误估计值和检验统计量。但不足的是，它们较大多传统方法更难以执行，而且每次执行多元回归，它都会带来不同的结果。

如果目标是要估计一个属于由 LISREL 或相似软件包所估计的模型种类的线性模型，则最大似然可能是首选方法。当前至少有 4 个统计套装软件可以完成这个任务，其中最知名的为 Amos。

如果你想要估计任何类型的非线性模型，则多重插补是正确的选择，有许多不同的方法可以执行多重插补。最广泛使用的方法是假设目标模型中的变

量有一多变量正态分布。通过涉及一个从数据值和参数随机抽取的重复回归的贝叶斯估计法来完成插补。当前有数个可用的套装软件可以完成这个任务。

其他做较少限制分布假设的多重插补方法目前正在发展中，但它们尚未达到理论上或计算上精细的改进水平，从而供正常化普遍使用。

也可以在数据不是随机缺失的假设下执行最大似然或多重插补，但要得到好的结果是很难处理的。这些方法对于缺失机制或有缺失数据的变量之分布的假设非常敏感。此外，没有方法检验这些假设。因此，最重要的必要条件是良好的关于缺失数据的产生机制的先验知识。且估计不可忽略的模型皆应伴随着一敏感性分析。

注释

[1] 证明很简单。我们想要估计，给定一预测变量的向量的 X 下，Y 的条件式分布 $f(Y|X)$。如果所有变量都被观察到令 $A=1$；否则，$A=0$。成列删除相当于估计 $f(Y \mid X, A=1)$。目标是要证明这个函数与 $f(Y|X)$一样。从条件式概率的定义而来，我们得到：

$$f(Y \mid X, A=1)=\frac{f(Y, X, A=1)}{f(X, A=1)}=\frac{\Pr(A=1 \mid Y, X) f(Y \mid X) f(X)}{\Pr(A=1 \mid X) f(X)}$$

假设 $\Pr(A=1 \mid Y, X)=\Pr(A=1 \mid X)$，也就是有数据呈现于所有变量上的概率不取决于 Y，但可能取决于 X 中的任何变量。接着：

$$f(Y \mid X, A=1)=f(Y \mid X)$$

注意这个结果可用于任何回归程序，不仅仅是用线性回归而已。

[2] 甚至当缺失数据的概率取决于 X 和 Y 两者时，有些情况下成列删除是没有问题的。令 $p(Y, X)$ 为一回归模型中一个或多个变量缺失数据的概率，以二分的因变量和一个自变量 X 的向量所表示的函数。如果该概率可以被因子化为 $p(Y, X)=f(Y)g(Y)$，则使用成列删除的 logistic 回归斜率为真实系数的一致的估计值(Glynn, 1985)。

[3] Glasser(1964)导出相当容易执行的公式，但是只当自变量和缺失数据形态在样本和样本间为“固定”时才有效，而这对于实际应用是一个不太可能的条件。Van Praag, Dijkstra 和 Van Velzen (1985)的公式更普遍地适用，但需要超出协方差矩阵中给定的信息：高阶动差和所有 4 个变量组的可得个案数目。

[4] 虽然当数据为确实缺失时虚拟变量调整方法是明显不能接受的，但它可能仍适用于未观察到的数据仅仅为不存在的个案的情况。例如，已婚受访者可能被要求评价其婚姻的质量，但该问题对于未婚的受访者就没有意义。假定我们有一个对已婚伴侣的线性方程及另一个对未婚伴侣的方程。已婚(伴侣之)方程等同于未婚(伴侣之)方程除了它具有(a)一个对应婚姻品质对因变量效应的项目以及(b)一个不同的截距。证明在这个情况下虚拟变量调整方法产生最适的估计值是容易的。

[5] 当使用条件式平均数插补时，Schafer 和 Schenker(2000)提出了一个(可)得到一致的标准误估计值(的方法)。他们宣称：在适当的条件下，他们的方法可以产生更精确的估计值，且较多重插补需要较少的运算努力。

[6] 在多变量正态假设下的最大似然，对于任何具有限四阶动差的多变量分布，产生平均数和协方差矩阵的一致的估计值(Little & Smith, 1987)。

[7] 这个变量为原始数据组中两个变量“房间支出”和“伙食变量”的总和。

[8] 对于二步骤方法，也可使用 Brown 和 Arminger(1995)叙述过的“三明治”公式得到标准误估计值。

[9] 在二变量正态性假设下估计的标准误，对于这个例子是适当的，因为数据从一个二变量正态分布抽取而来。公式为：

$$S.E.(r)=\frac{1-r^2}{\sqrt{n}}$$

虽然该样本相关系数不是正态分布的，但在这个例子中大样本数目应该确保密切近似反映正态性。因此，这些标准误可以适当地被用来建构置信区间。

[10] 对于数据扩增法，标准的不提供信息的先验分布(Schafer, 1997)为 Jeffreys 先验分布，写做 $|\sum|^{-(p+1)/2}$，其中Σ为协方差矩阵，p 为变量数目。

[11] 一个得到如此过度分散的分布的方法为使用一个自举的方法。例如，从原始数据组中有替代地取 5 个不同随机样本，并计算这些样本中的每一个 EM 估计值。EM 估计值接着可被用来作为 5 个平行系列中每一个的起始值。

[12] 一个执行这个过程的简单的方法是，将(0, 1)区间取比例于婚姻状态的每一个类别的概率的长度，分成 5 个次区间。对于随机数字落入的次区间指派相对应的婚姻状态。

[13] 事实上，自由度等于 L 的阶，通常但不总是为 r。

[14] 9 个地区分类如下：东部(新英格兰、中亚特兰大)，566 个个案；中部(东北中央、西南中央)，715 个个案；南部(南亚特兰大、东南中央、西南中央)，1095 个个案；西部(山区、太平洋)，616 个个案。

[15] 回归分析以 SAS 的 GLM 程序执行，使用 ABSORB 述句来处理固定效应。

[16] 产生 10 个数据组，每组 30 次反复。在插补后，插补值如有必要则被重新编码以保留原始变量可允许的值。

[17] 明确地，额外的变量为 $\lambda(ax_i)$，其中 a 为从 probit 模型估计系数的行向量。函数$\lambda(z)$，为反 Mills 函数，被定义为 $\phi(z)/\Phi(z)$，其中 $\phi(z)$为密度函数而$\Phi(z)$为累计分布函数，这两者都用于一个标准正态变量。

[18] 在 Rubin(1987)考虑的模型中，参数的条件式先验分布被指明缺失数据形态，在给定参数于完整数据形态的条件下。这里的实例中，我仅假设缺失个案的条件式平均数为完整个案的平均数的某个倍数。此外，缺失个案的条件式方差可以被允许大于那些完整个案的(方差)以弥补较大的不确定性。

参考文献

ANGRESTI, A. , and FINAY, B. (1977). Statistical methods for the social sciences. *Engle-wood Cliffs*, NJ: Prentice-Hall.

ALLISON, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg(Ed.), Sociological methodology 1987(pp. 71—103). Washington, DC: American Sociological Association.

ALLISON, P. D. (2000). Multiple imputation for missing data. *Sociological Methods & Research*, *28*, 301—309.

BARNARD, J. , and RUBIN, D. R. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*, 948—955.

BEALE, E. M. L. , and LITTLE, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, *Series B*, *37*, 129—145.

BROWNE, M. W. , and ARMINGER, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. C. Clogg, and M. E. Sobel(Eds.), *Hand Book of statistical modeling for the social and behavioral sciences*(pp. 185—249). New York: Plenum.

COHEN, J. , and COHEN, P. (1985). Applied multiple regression and correlation analysis for the behavioral sciences(2^{nd} ed.), *Hillsdale*, NJ: Erlbaum.

DAVIS, J. A. , and SMITH, T. W. (1997). General social surveys, 1972—1996. Chicago, IL: National Opinion Research Center(producer); Ann Arbor, MI: Interuniversity Consortium for Political and Social Research(distributor).

DEMPSER, A. P. , LAIRD, N. M. , and RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B*, *39*, 1—38.

FUCHS, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data, *Journal of the American Statistical Association*, *77*, 270—278.

GLASSER, M. (1964). Linear regression analysis with missing observations among the independent variables, *Journal of the American Statistical Association*, *59*, 834—844.

GLYNN, R. (1985). *Regression estimates when nonresponse depends on the outcome variables*. Unpublished doctoral dissertation, School of Public Health, Harvard University.

GOURIEROUX, C. , and MONFORT, A. (1981). On the problem of missing data in linear models. *Review of Economic Studies*, *48*, 579—586.

GREENE, W. H. (2000). *Econometric analysis*(4^{th} ed.), Englewood Cliffs, NJ: Prentice-Hall.

HAITOVSKY, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society*, *Series B*, *30*, 67—82.

HECKMAN, J. J. (1976). The common structure of statistical models of truncated, sample selection and limited dependent variables, and a simple estimator of such models. *Annals of Economic and Social Measurement*, *5*, 475—492.

IVERSEN, G. (1985). Bayesian statistical inference. Sage University Papers Series on Quantitative Applications in Social Science, 07—43. Thousand Oaks, CA: Sage.

JONES, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*, 222—230.

KIM, J.-O., and CURRY, J. (1977). The treatment of missing data in multivariate analysis, *Sociological Methods & Research*, *6*, 215—240.

KING, G., HONAKER, J., JOSEPH, A., and SCHEVE, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, *95*, 49—69. Available at http://gking.harvard.edu/stats.shtml.

LANDERMAN, L. R., LAND, K. C., and PIEPER, C. F. (1007). An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods & Research*, *26*, 3—33.

LI, K. H., MENG, X. L., RAGHUNATHAN, T. E., and RUBIN, D. B. (1991). Significance levels from repeated p-values and multiply imputed data. *Statisica Sinica*, *1*, 65—92.

LITTLE, R. J. A. (1988). Missing data in large surveys(with discussion). *Journal of Business and Economic Statistics*, *6*, 287—301.

LITTLE, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*, 1217—1237.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125—134.

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, *81*, 471—483.

LITTLE, R. J. A., and RUBIN, D. B. (1987). *Statisical analysis with missing data*. New York: Wiley.

LITTLE, R. J. A., and SMITH, P. J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, *82*, 58—68.

MARINI, M. M. OLSEN, A. R., and RUBIN, D. (1979). *Maximum likelihood estimation in panel studies with missing data*. In K. F. Schuessler(ed.), *Sociological methodology 1980* (pp. 314—357). San Francisco: Jossey-Bass.

McCULLAGH, P. (1980). Regression models for ordinal data(with discussion). *Journal of the Royal Statistical Society*, *Series B*, *42*, 109—142.

McLANCHLAN, G. J., and KRISHNAN, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

MOSSEY, J. J., KNOTT, K., and CRAIK, R. (1990). Effects of persistent depressive symptoms on hip fracture recovery. *Journal of Gerontology: Medical Sciences*, *45*, M163—168.

MUTHÉN, B., KAPLAN, K., and HOLLIS, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *42*, 431—462.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOOEWYK, J., and SOLENBERGER, P. (1999). *A multivariate technique for multiply imputing missing values using a sequence of regression models*. Unpublished manuscript. Contact teraghu@umich.edu.

ROBINS, J. M., and WANG, N. (2000). Inference for imputation estimators. *Biometrika*, *87*, 113—124.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581—592.

RUBIN, D. B. (1987). *Multiple Imputation or Nonresponse in Surveys*. New York: Wiley.

RUBIN, D. B., and SCHENKER, N. (1991). Multiple imputation in health-car databases: An overview and some applications. *Statistics in Medicine*, *10*, 585—598.

SCHAFER, J. L. (1997). *Analysis of incomplete Multivariate Data*. London: Chapman & Hall.

SCHAFER, J. L., and SCHENKER, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, *95*, 144—154.

SCHENKER, N., and TAYLOR, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, *22*, 425—446.

STOLZENBERG, R. M., and RELLES, D. A. (1990). Theory testing in a work of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods & Research*, *18*, 395—415.

STOLZENBERG, R. M., and RELLES, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, *62*, 494—507.

VACH, W. (1994). *Logistic Regression with Missing Values in the Covariates*. New York: Springer-Verlag.

VACH, W., and BLETTNER, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, *134*, 895—907.

VAN BUUREN, S., and OUDSHOORN, K. (1999). Flexible multiple mputation by MICE. Report TNO-PG 99. 054, TNO Prevention and Health, Leiden. Available at http://www.multiple-imputation.com/

VAN PRAAG, B. M. S., DIJKSTRA, T. K., and VAN VELZEN, J. (1985). Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem. *Psychometrika*, *50*, 25—36.

WANG, N., and ROBINS, J. M. (1988). Large sample inference in parametric multiple imputation. *Biometrika*, *85*, 935—948.

WINSHIP, C., and RADBILL, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, *23*, 230—257.

译名对照表

aspect	面向
autocorrelation	自相关
available case analysis	可得个案分析
Bayesian bootstrap method	贝叶斯自举法
Bayesian posterior distribution	贝叶斯后验分布
casewise deletion	个案删除
categorical independent variables	类别自变量
cell	单元格
censor	删截
chi-square	卡方
column vector	栏向量
complete-case missing-rariable restrictions	完全个来缺失变量限制
conditional distribution	条件式分布
conditional mean	条件平均数
conditional mean imputiion	条件式平均数插补
confidence interval	置信区间
contingency table	列联表
correlation	相关性
covariance	协方差
covariate	协变量
Cox regression	Cox 回归
Cox proportional hazards model	Cox 比例风险模型
cross-product ratio	交叉相乘比
cumulative logit model	累进的 logit 模型
cut-off point	截略点
data	数据
data augmentation(DA)	数据扩增法
default	默认值
denominator degrees of freedom(d. d. f)	分母自由度
departure	偏离
dependence	依赖性/相关性
dependent variable	因变量
diagnostic statistics	诊断统计量
dichotomy	二分法
dispersion	分散

disporportionate stratified sampling	非比例分层化抽样
distribution	分布
dummy variable adjustment	虚拟变量调整
effect	效应
efficient	有效的
error	误差
estimate	估计值
estimator	估计量
expectation-maximization(EM)	期望最大化
exploratory analysis	探索分析
factor analysis	因子分析
factor loadings	因子载荷
failure-time regression	失效时间回归
fit	拟合
fraction missing information	缺失信息比
frequency	频数
generalized linear model	广义线性模型
Heckman's selectivity bias model	Heckman 的选择性误差模型
heteroscedasticity	异方差性
higer-order moments	高阶动差
homoscedasticity	同方差性
Hot deck method	热卡方法
imputation	插补
independent variable	自变量
intended model	预设模型
iteration	迭代
joint distribution	联合分布
lags	滞后值
latent variable	潜变量
least square	最小二乘
logistic regression	logistic 回归
log-linear analysis	对数线性分析
listwise deletion	成列删除
marginal distribution	边际分布
marginal mean imputation	边际平均数插补
Markov chain Monte Carlo(MCMC)	马尔可夫链蒙特卡尔

maximum likelihood(ML)	最大似然
mean substitution	平均值替换
metric	度量标准
missing at random(MAR)	随机缺失的
missing completely at random(MCAR)	完全随机缺失的
missing data mechanism	缺失数据机制
missing data-generating mechanism	缺失数据产生机制
missing indicator method	缺失指标方法
monotonic	单调的
multinomial	多项的/多峰的
multiple imputation(MI)	多重插补
multiple imputation by chained equations(MICE)	基于链式方程的多重插补
normal distribution	正态分布
null hypothesis	零假设
observed at random	随机观察的
odds	比例
overall likelihood	总体似然
overidentified model	过度识别模型
normal distribution	正态分布
p value	P 值
pairwise deletion	成对删除
panel study	面板研究
patterns	形态
poisson regression	泊松回归
population	总体
positive definite	正定的
predictive mean matching	预测均数匹配
predictor	预测(量)
probability	概率
probability density function(p. d. f.)	概率密度函数
propensity score	倾向评分
proportional odds assumption	成比例发生比假设
portion	比率
prior distribution	先验分布
ratio	比、比率
residual	残差

robust	稳健的
root mean squared error(RMSE)	均方根差
row rector	列向量
sample size	样本数
sampling variablilty	抽样变异
simulation study	模拟研究
simultaneous equations	联立方程
skewed	偏(态)的
skewness	偏度
standard deviation	标准差
standard error	标准误
stratified sampling	分层化抽样
structural equation modeling	结构方程模型
summary statistics	描述性统计
systematic error	系统性误差
t statistic	t 统计量
valid	有效的
validity	效度
variance	方差

社会网络分析

作 者 简 介

戴维·诺克(David Knoke)

美国明尼苏达大学双城分校教授,曾于 1972 年获密歇根大学博士学位。和 Kuklinski 合著《网络分析》(*Network Analysis*)(1982)一书,出版 15 部著作,发表 100 多篇论文和论著章节,其主要研究领域为组织、网络、政治学和社会统计学。他是多项国家科学基金项目主持人,其项目包括自愿者组织、国家政策领域的游说团体以及多种不同机构的组织调查,最近的一个项目涉及全球信息产业,关注跨国公司间战略合作关系的发展。

杨松(Song Yang)

美国阿肯萨斯大学教授,曾于 2002 年获明尼苏达大学博士学位。著有《美国组织内工作培训项目调查》(*Investigating Job Training Programs in U. S. Organizations*)(2003)一书,发表有关社会网络分析的论文 15 篇(含著作章节)。

译 者 简 介

李 兰

香港科技大学社会学在读硕士研究生。获哈尔滨工业大学机电学院本科学位、北京大学人口研究所硕士学位。目前的主要研究兴趣为社会研究方法、社会分层、教育社会学及人口学等。

序

社会网络分析从一开始就是一种跨学科的研究方法。与其他源于单一学科的研究方法如多维度测量法不同，网络分析的起源至少能追溯到三种学科：心理学、人类学和社会学。在心理学领域，社会计量法杂志（*Seciometry*）（1937）的创始人、心理医生 Jacob Moreno 研究出同名法。同名法是用来测量社会关系的，以便更好地研究社会结构和心理健康的关系。人类学家 W. Lloyd Warner 和心理学家 Elton Mayo 主持研究芝加哥西部电力公司的霍桑电厂时，Roethlisberger 和 Dickson 利用这项研究于 1939 年发表了有关触排布线室的数据，即使到今天，这种分析仍然不失为一项非常出色的研究。关于这项数据，George Homans（1950），Ronald Breiger 和他的同事（1975）利用社会学方法重新进行了分析，以使更多人有机会了解这项数据。包括 Harrison White 和 Mar Granovetter 在内的社会学家于 20 世纪六七十年代对推动理解社会网络、完善社会网络分析方法，亦作出了重要的贡献。

本书第一版于 1982 年出版，当时书名为《网络分析》，作者为网络分析文献专家 David Knoke 和 James Kuklinski。从那时起，社会网络分析大步向前发展，影响了社会科学的很多领域。Linton Freeman 于 1979 年创办了《社会网络：一本结构分析的国际杂志》，成为该领域的旗舰杂志。许多国内或国际的社会调查都按常规把社会网络分析内容当做调查的一部分，甚至多次重复调查这些问题。例如，美国 1985 年的综合社会调查（GSS）收集的网络调查数据被广泛用于各种社会科学文献中。从 1994 年到 2002 年，美国做的三次全国青少年健康纵向调查都收集了社会网络方面的数据。在新千年里，欧洲的研究者首次做了欧洲生活质量的调查，其内容包括家庭、工作和生活网络等。近年来，有关社会网络分析的计算机软件如雨后春笋般涌现，而且都可以在网上免费获得。社会网络研究人员现在至少有 50 多种程序可以选用，例如用于社会网络分析、社会计量和序列分析的 Agna，以及 Tom Snijder 用于分析 0/1 矩阵的程序 ZO。

面对过去二十多年的这些发展，本书的第一版显然已经过时，亟须更新再版。David Knoke 及合著者 Song Yang 担起此任，重写了本书，现命名为《社会网络分析》。和第一版相同，第二版开始介绍了一些基本概念和数据收集问题，然后开始介绍网络分析的方法

和模型。然而，即使在基本概念这部分，新版本也添加了很多重要内容，如讨论网络分析的三个主要的潜在假设。而且，这本书还反映了社会网络分析的许多最新进展，例如，更多地强调图形及 p* 模型这样的高级方法。网络研究人员及学生将会发现，本书为他们探究社会网络数据提供了许多极好的指导。

——廖福挺

第 1 章 | 社会网络分析简介

近几十年来，社会网络越来越受到公众和学界的青睐。John Guare 的电影作品《六度分隔》(*Six Degree of Separation*)使得社会心理学家 Stanley Milgrom 的小世界理论众所周知。小世界理论指的是，任何两个人都可以通过几个紧密连接的中介产生关联。该理论的另一种流行文化表现形式是“凯文 · 培根游戏”，即影迷只需小小的努力，就可以通过演员的搭档及搭档的搭档让两个成名演员通过几个搭档联系起来。现在，在高中生和大学生中，MySpace 和 Facebook 等在线网络交流如雨后春笋般涌现；与此同时，商务部门也不断建议利用网络来寻求新的商业利润和职业发展机遇(Cross, Liedtka & Weiss, 2005)。

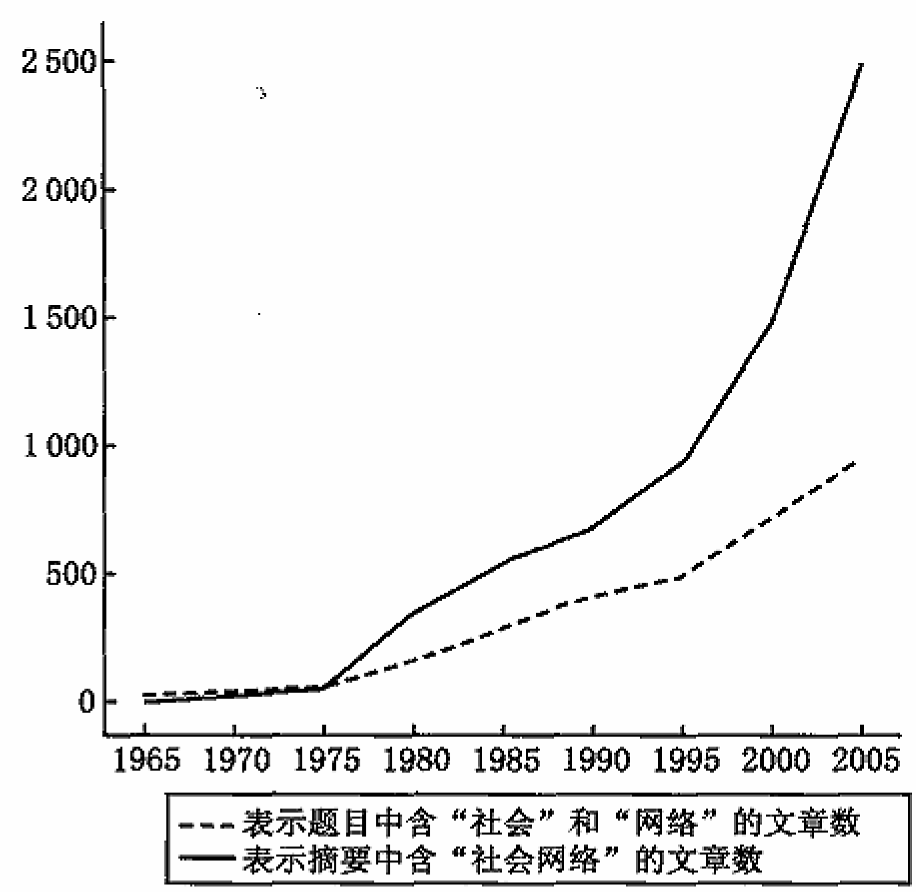


图 1.1　社会科学文献中有关“社会网络”研究文章的增长情况

如图 1.1 所示，在过去 30 年里，关键词为"社会网络"的社会科学出版物成指数增长，且毫无变缓的趋势。类似地，网络研究也迅速被应用到自然现象及数学和物理领域中复杂的社会系统中(Barabási, 2002; Newman, 2003; Watts, 2003)。网络分析已然成为一种制度化的跨学科的视角，它的基本概念和测量方法已为多个领域的研究人员所熟知，如社会学、人类学、经济学、组织研究、工商管理、公共卫生、信息科学、生物学、复杂理论和混沌理论。

同时，可用于网络分析的计算机软件也相应地快速增长，如 UCINET, Pajek, NetMiner, STRUCTURE, MultiNet，和 StOCNET 等通用程序被广泛应用于常规的网络数据分析(详细介绍请参见 Huisman & van Duijn, 2005)。同时，也出现了许多有特殊用途的程序来处理识别亚群体(Frank, 2003)、研究隐性人群(Frank & Snijders, 1994)、统计检验(Tsuji, 1997)和计算赋值图中的最优连接(Yang & Hexmoor, 2004)等特殊问题。UCINET 可能是最受欢迎、使用范围最广的软件包，版本不断更新，为许多网络分析方法的实现提供了综合的解决方案。它还提供了各种文件输入和输出、数据操作、网络工具(MDS，聚类和统计分析)和网络方法(位置、角色、亚群体、块模型和自我中心分析法)等级联模块。基于上述理由，本书将使用 UCINET 软件来分析不同的网络实例、阐释有关的网络分析。

本书是 1982 年 Knoke 和 Kuklinski 合著的《网络分析》(第一版)的全面修订版。除了综述有关基本方法的论题外，我们还囊括了过去四分之一个世纪的研究新进展。但是，由于篇幅的局限，本书不可能像 Wasserman 和 Faust (1994)那样进行全面介绍，但与 Degenne 和 Forse(1999)、Scott(2000)的入门书籍相比，本书介绍了更多的高级方法。本书为进一步学习 Carrington、Scott 和 Wasserman(2005)介绍的更深入的高级方法打下了一定的基础。

本书主要应用于教学，针对的读者主要是社会科学领域的研究生和专业人士，其中包括社会学、工商管理、人类学、经济学、公共卫生和人力资源等。这本书可作为研究生课程的教学参考书，可用于专业学会或暑期研修班的研讨会，当然也可用于网络研究的自学。大学高年级学生和研究生如果对社会网络分析感兴趣，也可通过阅读本书在社会网络分析技巧和学术志向方面获得更高的

起点。很多专业人员在做社会网络研究时会面临许多挑战，例如如何设计一个社会网络项目，网络数据收集过程中可能出现的各种细节和问题，以及分析社会网络数据时其他可以替代的技术等。社会网络学者可以把本书作为一本简单实用的复习用书或参考书。

在本书中，我们常常利用实际的社会网络研究问题来解释相关的概念和方法，引用的例子包括孩童游戏小组、组织、社区和跨国集团。我们尽量避免使用复杂的专业术语，而是用简洁的语言来进行精确介绍。本书共包括 4 个部分。第 2 章介绍有关关系数据和网络的概念，这些概念可应用于一系列的观察单元、分析层次和测量类型。本章比较了关系的内涵和构成，区分了自我中心网络和全幅网络。结构方法强调网络分析揭示经验交流背后更深层模式的价值。第 3 章主要关注数据收集问题：边界界定，全幅网络和网络抽样，提名法和提名清单，偏差和可靠性，自我中心网络和全幅网络的关系测量，以及缺失数据。第 4 章讨论网络分析的基本方法，包括图表法和矩阵法；中心度、声望和权力；小团体、内聚和连接；结构等价；视觉演示、聚类和多维标度；块模型。第 5 章介绍了一些更高级的网络分析方法，包括位置和角色；自形等价、同构等价和正规等价；Logit 模型(p*)；隶属网络和点阵分析。读者如果希望通过分析书中的数据，以便学习如何开始进行网络分析，可以在本书作者 Knoke 的主页 http://www.soc.umn.edu/~knoke/上找到书中所有实例的数据集和相应的计算机指令的链接，并且可以下载文件压缩包。

第 2 章 | 网络基础

本章讨论社会网络分析方法的基本概念。我们使用学者普遍使用的术语和定义，但在存有分歧的情况下，遵从社会学惯例。我们在解释这些基本概念的同时，说明这些概念的文献出处。感兴趣的读者可以通过这些文献，更好地理解网络分析方法是如何应用到实际问题的研究之中的。

为了更好地理解我们所展示的社会网络分析视角的独特性，这里将基于变量的个体分析方法与之进行一下比较。社会科学的许多领域，甚至社会科学的绝大部分领域，都假定行为人作出决定和采取行动时，不考虑其他行为人的行为。无论是基于效用最大化的理性计算来分析，还是基于因果关系的动因推理来分析，这些解释仅仅主要考虑了个人特性，而忽略了社会行为人所嵌入的更广泛的互动背景。与此不同的是，网络分析法明确假定，行为人参与到社会系统之中，而这个系统将他们与其他行为人联系在一起；行为人之间的关系对各自的行为存在着重要影响。网络分析理论和方法的关键，是识别、测量和检验关于行为人之间关系的结构形式和实质内涵的假设。这种对结构—关系研究的独特重心，将社会网络分析法与迄今依然主宰大部分社会科学的以变量为中心的个体分析的传统区分开来。

基本假定

网络分析的视角强调把结构关系作为关键的导向原则，其中社会结构的组成是“有形实体之间关系模式的规律性，而不是抽象的规范和价值之间的协调，也不是对这些有形实体的特征进行的分类”(White, Boorman & Breiger,

1976:733—734)。实体可以是单个的自然人,小的群体,组织,甚至是国家。通常连接这一系列实体的关系模式包括宏观的社会背景或者整体的结构,这些背景和结构影响他们的认知、信念、决定和行动。网络分析的首要目标是精确测量和展现这些结构关系,并解释他们发生的原因及产生的结果。

社会网络分析的重要性依赖于有关关系模式及其效用的三个基本假设。首先,对于理解所观察的行为,结构性关系要比年龄、性别、价值和意识形态等个体特征更为重要。例如,在解释15世纪的佛罗伦萨哪些精英家族支持美第奇(Medici)或寡头政治集团的统治时,更多地依赖这些统治精英之间的经济、赞助和婚姻关系,而较少依赖他们的阶级和社会地位等个人特征(Padgett & Ansell, 1993)。在实体参与的不同社会情境中,他们的许多特征保持不变(如不论在家、在工作场所还是在教堂,一个人的性别、种族和教育程度都是不变的)。但是,具体的结构关系仅存于特定的时空;当情境发生变化时,这种特定的结构关系将消失或暂时中止(例如,学校之外就不存在老师和学生之间的关系;当一个人丧偶或离婚时,婚姻关系将不存在)。一个在办公室做勤杂工的女性,在工作场所可能不需要什么主动性,但她却可能是邻居关系协会中强有力的领导者,或者是个敢作敢为的PTA① 成员。这种行为差异很难用性别、年龄和社会地位等不变的特征来解释;但是,当意识到人与人之间的结构关系可以随着社会情境的改变而变化时,这种差异就很好理解了。这种结构关系的分析方法备受网络分析者的欢迎;而与之形成鲜明对比的是诸多的互争高低的实体论方法,这些方法将静态的"事物—概念"作为自己的主要分析单元,包括:本质、自我行为、基于规范的一致、理性选择、以变量为中心和社会认同途径等等(Emirbayer, 1997)。网络分析认为,除了个体特征,社会实体间的已具有模式的关系也影响社会实体。根据这样的推测,网络分析宣称,可以为分析社会行为的原因提供更加全面的理论和实证解释。

第二,社会网络通过实体之间的关系所建立的各种结构机制影响实体的观念、信仰和行动。直接关系和更密集的互动关系使得实体获得更好的信息、更

① 即家长—教师关系协会。——译者注

多的认知、更强的敏感性,以影响他人或受他人的影响。间接关系通过媒介(例如常见的经纪人,他们为客户牵线)也会激发新想法,或获得一些潜在的有用资源,而这些资源可能需要通过与他人交易才能获得。例如,相对于较弱、较疏远的社会关系,求职人员一般很难通过更密切的社会关系获得有价值的信息,因为关系更亲近的人已经共享或传播了相同的信息(Granovetter, 1973)。关系结构提供了各种复杂路径以协助或阻止人群中知识、流言蜚语的传播。社会结构配置上的不同,解释了芬兰有机耕作的普及(Nyblom, Borgatti, Roslakka & Salo, 2003)和美国城市中性传染疾病的传播(Jolly, Muth, Wylie & Potteral, 2001)。个人的身心健康相当大程度上依赖于社会支持网络——比如,当你沮丧的时候,是否有人给你送上一份心灵鸡汤(Cohen et al., 1998)。结构关系对维持团体的凝聚力和团结至关重要,但也可能强化对外部群体的偏见和冲突。竞争与合作关系有效实现了跨国公司内部有计划的组织变迁(Tenkasi & Chesmore, 2003)、社会运动中集体行动的动员(Diani & McAdam, 2003),以及从事贩毒、非法移民、走私和恐怖活动等"黑色网络"的运转(Raab & Milward, 2003)。通过把信息和资源传递到特定的结构位置,网络有助于激发兴趣、分享认同、促进共同的价值和规范。网络分析者试图揭示社会关系如何影响社会实体的理论机制,并发掘出特定机制在具体情境中运行可能需要的条件。

网络分析的第三个潜在假设是应将结构关系视为动态过程。这个假设认识到网络不是静态结构,而是通过个人、群体或组织等实体内部的互动不断改变的。当这些实体利用有关网络的知识获取优势时,他们或有意或无意地改变了他们所处的关系结构。例如,在 Kenis 和 Knoke(2002)有关组织领域网络变迁的理论中,就认为先前的沟通结构会影响后续的战略联盟抉择。相应地,这些组织间的关系改变了信息的流动,从而为将来的联盟提供了更多的机会或约束。这些动态过程验证了社会行动理论中更为普遍的"从微观到宏观的问题"(Coleman, 1986)。关键问题是大规模的系统性变化如何从个人的综合偏好和目的性行动中显露出来。由于网络分析同时涵盖实体和结构,因此它为微观层面选择的变化和宏观层面结构的变迁之间的连接提供了有效的分析概念和方法。不幸的是,跨层动态过程的实证网络分析仍只是一个理想的目标,并没有

成为一个普遍的做法(Emirbayer, 1997:305)。但是,随着纵向网络数据的逐渐增多,对网络变迁进行抽象和调查的新方法也将出现(Faust & Skvoretz, 2002; Snijders, Steglich, Schweinberger & Huisman, 2007)。

行为人和关系

任何社会网络中两个不可或缺的元素是行为人和关系。它们的结合共同构成了我们下节要介绍的社会网络。行为人可以是自然人等个体,或非正式群体、正式组织等集体。个体行为人的例子包括操场上的小孩,参加舞会的高中生,公司里工作团队的职员,疗养院的员工和住客,秘密基层组织中行动的恐怖分子等。集体行为人可能是在一个行业中竞争的公司,筹集慈善基金的志愿者组织,在国会里拥有席位的政党,也可能是签署军事同盟协议的国家。有时网络行为人也包括一些混合类型,如一个组织领域包含有供货商、生产商、顾客和公共卫生的政府监管人员。

关系一般被定义为两个行为人或对点之间的特定接触、连接或联结。关系可以是定向的,即第一个行为人发起,第二个行为人接受(如劝告);关系也可以是非定向的,即存在互动(如交谈)。关系不是一个行为人的特征,而是在只有两个行为人保持联系时一种共同的对偶特征。在个体行为人以及集体行为人之间,会发生许多各种不同的关系,这些关系可展示网络结构,并解释它们的效应。比如,在个人之间交往的层面上,小孩之间交朋友、一起玩耍、打架、相互信任,同事之间一起工作、讨论、提意见、信任、背后说坏话和背叛。而在集体行为人之间,公司进行商品和服务的交易、相互交流、竞争、起诉、游说和合作。一个网络研究者应该测量哪一种关系,这取决于研究项目的特定目标。例如,社区网络研究可能调查各种不同的邻里活动,而银行网络研究则主要调查财务交易。当然,一些分析详细研究多类关系,如公司董事会董事之间的政治、社会、经济联系。下节我们将介绍关系内涵的一般分类。

很多社会研究无论是通过调查还是实验来收集数据,依旧依赖于测量和分析行为人的特征,并把行为人的特征作为分析单元。虽然特征和关系是研究社

会行为的不同概念方法，但是它们不能视为互斥的选择。相反，许多行为人的特征可以重新概念化为两个行为人之间的关系。因此，虽然一国每年的进出口额是这个国家经济的特征，但所有的任何两个国家间交易的商品和服务量却可以表示全球经济的交易网络结构。一个棒球队的总输赢纪录显示这个球队的表现，但每两个队的输赢比例可能揭示一个联盟权利结构的细微差别。朋友的数量显示了一个小孩的受欢迎程度（人气），但分析每一对朋友关系的选择就能揭示出重要的小团体和群体。关系反映出复杂社会系统中仅靠成员特征的求和或均值所不能描述的新维度。结构关系能影响个人行为和系统表现，且这些行为和表现不能还原为行为人的特征。例如，与其考虑行为人的教育、年龄、社会阶层、性别和种族，不如考虑行为人之间的交流和建议关系结构，这样能更好地解释技术创新的传播比例和模式（Cowan & Jonard, 2004; Valente, 1995）。类似地，和推测群体的易受感染性和抵抗力特征相比，熟知群体间的亲密关系结构更有利于防止像禽流感和艾滋病等传染性疾病的流行爆发（Jolly et al., 2001）。在此，一个非常重要的推论是：仅仅分析行为人的特征，会忽略社会行为的网络视角所提供的许多重要解释。

网络

社会网络是由一群行为人组成的结构，在这个结构中行为人通过一系列关系相连。社会网络的绝大多数定义都包括这两个基本组成部分。例如，“社会结构可以表述为网络——包括一系列节点（或社会系统的成员）和一系列描述他们之间关联性的关系”（Wellman & Berkowitz, 1988:4）。即使行为人都相同，不同类型的关系也可以形成不同网络。因此，一群雇员之间的朋友网络很可能与他们之间的建议寻求网络存在很大差异。网络行为人之间存在联系并不意味着所有行为人必须和其他行为人之间存在直接关系。实际上，有时直接的对偶关系并不多见。相反，网络分析会同时考虑目前存在的或不存在的关系，也可能会把关系的强度变化考虑进去。有形实体之间的实际关系构造，便确定了一个特定的网络结构——该网络的模式或形式。网络结构的形式变化

多端，既有所有行为人之间没有关联的孤立结构，也有所有行为人之间直接相关的饱和结构。一般来说，实际的网络结构处于独立网络结构和饱和网络结构之间，其中一些行为人比其他行为人拥有更广的网络。网络分析中的一个核心理论问题，便是解释不同结构的发生情况，以及在行为人层面来说明其与其他行为人之间关系的变化。在网络研究中，一个同样重要的实际任务是利用关系数据，发掘并精确展现这些结构。

John A. Barnes (1954)被认为是第一个使用社会网络这一术语的学者。他是一个人类学家，曾经研究挪威一个岛上居民之间的关系。Barnes 将社会互动视为“点的集合，其中一些点由线连接”，从而形成关系的“全部网络”(Barnes, 1954:43)。人际关系的非正式形式构成社会网络整体的一个部分。Barnes 利用了 Jacob Moreno(1934)的成果，后者在展示小孩对同班同学的爱憎关系研究中，率先使用了线和标示点的手绘社会关系图。本书将在第 3 章讨论如何表示有关社会网络关系的形象图示和数学矩阵。半个多世纪以来，社会网络这一理念和方法已经从人类学和社会学延伸到诸多学科，并被运用到诸多主流的理论和问题研究中。关于社会网络原理缘起及普及的历史性综述，可参见 Freeman (2004)，Scott(2000)及 Knox，Savage 和 Harvey(2006)的著作。

假如社会网络分析仅仅只是描述一群行为人是如何联系的一个概念性框架，那么它就不会激起社会研究者如此大的兴趣和努力。作为一套综合的理论概念和分析方法，社会网络分析所能提供的不仅仅是精确重述。由于社会网络结构影响到个体层次分析和系统层次分析，因此它被认为可以用来解释结构关系中的变化及其结果。J. Clyde Mitchell(1969:2)关于社会网络的定义，就特别强调了它们对结果的影响。他的定义如下：社会网络是“在特定人群中的一整套具体的联系；这种联系有一个额外特征，即这些联系作为一个整体可以用来解释其中人们的社会行为”。这一论述在本书的第一版中也得到了回应：“行为人之间的关系结构和个体行为人在网络中的定位，对于每个个体单元和整个系统来说，在行为、意识和态度方面都有重要的影响。”(Knoke & Kuklinksi, 1982:13)同样，Barry Wellman(1999:94)也写到：“在描述社会结构的基本模式，在解释这些模式对人们行为和态度的影响上，社会网络分析都能起到作用。”

研究设计的基本要素

对于研究者来说，构成测量和分析框架的社会网络研究设计的三个基本要素是：抽样单位，关系的形式和内容，数据分析层次。每一个网络研究项目都必须在开始田野工作之前对每个基本要素作出明确界定。这些设计因素的不同组合将生成各种不同的社会网络调查，这将都呈现在多个学科的研究文献中。

抽样单位　一项社会网络研究设计的第一步，即选择一个最相关的社会背景，并确定在此背景下的哪些组织涵盖了这些社会网络行为人。根据规模和复杂性大致的升序排列，构成抽样单位的6个基本单元，包括个体、群体(正式和非正式)、复杂的正式组织、阶级与阶层、社区，以及民族—国家。一些两阶段的研究设计包括一个高层次系统，在此系统中行为人构成其中的低层次实体。这种嵌套社会背景的普通例子包括公司与职员、学校与学生、市政机构与公务员、大学及其院系与教授等。

最早且至今仍然流行的社会网络研究项目一般选择小规模的社会背景——比如课堂、办公室、工厂、帮会、社交俱乐部、学校、村庄、人为组织的试验群体，在此将其中的个体成员作为社会关系的行为人，而他们的关系正好构成了受调查的网络。最新的例子有：加州少年课堂中与欺负和被欺负相关的朋友关系模式(Mouttapa, Valente & Gallaher, 2004)，两家冶金工厂全方位质量管理方案的实现(Harrison, Laplante & St-Cyr, 2001)，以及马达加斯加一个村落中家庭生育控制方式的社区分布情况(Stoebenau & Valente, 2003)。在精确描述成员资格、全面观察群体、易于通过上层权威打通关系等方面，小规模背景都有明显优势。较大规模的环境往往都有互相渗透的或者模糊不清的边界，但是社会网络分析法并不因其内在特性而不适宜将其概念与方法运用于这类环境。最近的研究范例有：某一中西部城市青少年的恋爱与性行为关系网络(Bearman, Moody & Stovel, 2004)，学院派经济学家的合著关系(Fafchamps, van der Leij & Goyal, 2006)。

关系的形式和内容　除了选择一个恰当的抽样单元之外，社会网络研究者

还必须决定收集哪种具体社会关系的数据。行为人之间的关系既有形式又有内容，这种两分法概念是由 Georg Simmel（1908）在分析社会交往时提出的。形式和内容这二者在实践中是不可分的，只是为了分析的方便而将二者区分开来。内容是指社会交往中人们的兴趣、目的、压力或者动力，而形式则是由此而将具体内容转化为社会现实的互动模式。Simmel 认为，社会学的任务就是发掘确认社会交往有限的几种形式：交际、超越、臣服、竞争、冲突、合作、统一等。这些交往形式是在广泛的具体环境、社会制度和历史背景下发生的。同一种具体形式在内容上可以千差万别。例如，上级和下级这一基本关系形式，可以在政府、军队、商业、宗教、体育运动和文化机构等不同组织中展现。相反，不同的内容——比如追逐经济利益和政治权力——也可以通过竞争与合作的形式展现。

内容与形式这一二分法概念也运用到社会网络分析中。关系形式是行为人关系的一个特征，它独立存在于所有的具体内容之中。有两种基本的关系形式：(1)双方行为人之间互动关系的强度、频率或者力度；(2)每对成员之间关系的指向，即是否为 0，或者非对称的，或者双向的选择。关系内容指作为关系发生原因的那些事物（Burt，1983：36）。这种实质性内容是由研究者建立的一种可分析构造，意在从行为人主体视角获取某一关系的意义。当人们在某一社会背景中被问及“你的亲密朋友、一般朋友和认识的人”时，这一问题的内容就是指“友情”。这一询问的结果得依每一行为人是如何定义这三个词所代表范畴的意思而定，另外还得依行为人是如何根据不同人际互动的以往经验来区别对方而定。很明显，人们在对友情这一标签及表示亲密或疏远的行为的理解上，可能会有很大差别。

关系内容或者说连接模式的选择，大部分是由项目的理论焦点和研究目标来决定的。因此，一项关于保健关系网络的研究就会调查人们有关健康事项信息和建议的来源，而一项关于政治关系网络的研究也许会让他们列出平时经常和他们一起讨论或参与政治事务的人。在一些实证问题的研究中，也许不仅仅某一个独特的可分析关系内容需要调查，而且其他两三种甚至更多的连接模式（即复杂关系网络）同时也可能需要进行测量和分析。一项关于

德国、日本与美国政府劳工政策游说机构的比较研究，要求受访人报告他们的独立政策信息、政治支持、与他人的信息互换关系等(Knoke，Pappi，Broadbent & Tsujinaka，1996)。我们将在第4章中利用其中的一些数据来阐述有关网络分析的方法。

虽然在Simmel看来，对于将关系内容的细微差别进行概念化和操作化，研究者们几乎有着无限的想象力，但是在这里我们还是提供一个小小的基本内容范式：

(1) 交换关系：行为人互换有关实物或象征性物质的控制，比如赠送礼物或者买卖商品。

(2) 沟通关系：通过行为人之间的联络渠道，双方交换有关信息。

(3) 交叉关系：行为人在两个或多个社会组织中有成员资格，例如公司董事会成员同时拥有其他身份。

(4) 利用关系：行为人之间互相联系，是为了获得有价值的物品、服务或者信息，比如一份工作、一个政策建议或者一个加入某项社会行动的机会。

(5) 情感关系：也许最常接受调查的社会网络关系就是行为人对他人的感动、崇拜、尊敬、厌恶或者敌视等情绪的表达。

(6) 权力关系：这类联系通常发生在正式的科层组织中，表明行为人对于发布和遵守命令的权利和义务。

(7) 亲属关系：这类血缘和婚姻的联系反映家庭角色内的关系。

社会网络分析对于各种各样的关系内容是如何连接的仍然认识不足，因为对于关系内容领域结构的考察仍然研究得不够。Ronald Burt(1983)在关于调查对象是如何认识关系内容的研究中，发现在抽样调查北加州人的33个问题中，存在实质性的混淆、重复和可替代等情况。他认为，仅仅只需5个关键问题就足以涵盖有关友情、相识、工作、亲属和亲密关系这几种关系内容的主要结构。然而，对于在各种不同网络环境中普遍使用的关系术语和标签，人们所赋予其意义的相似性和差异性，我们还需要更多的研究。一张关于关系内容领域间的结构性连接的认知图，可以帮助研究者准确有效地选择最适合理论和实践关注点的具体关系内容。

数据分析层次　在选择了抽样单元、关系的形式和内容之后，研究者对于他们在社会网络研究项目中所搜集到的数据，在哪一个层次上进行结构分析，有多个可替代性选择。本书第 3 章和第 4 章会详细论述有关合适的测量和分析方法。这里我们将首先概述 4 种概念独立的分析层次，调查者可能会从中选择决定集中分析。

最简单层次是自我中心网络。这一社会网络由一个行为主体和其他行为客体组成，其中行为主体与其他行为客体发生直接联系，而其他行为客体之间也存在直接联系。这一层面也被称为行为主体的"第一关系圈"；相对于第一圈，所有行为客体的客体构成了第二关系圈和更高层次的关系圈，以此类推。假如一个网络中有 N 个行为人，那么自我中心网络则会有 N 个分析单元。相应地，每一主体行为人可以通过其与其他行为客体联系的数量、强度和其他特征加以描述，例如报酬与付出关系的比例，或者行为客体之间联系的密度。最近一项关于毒品成瘾者个人健康风险的自我中心网络研究发现，波多黎各的瘾君子在与陌生人或认识的人发生性行为中，要比纽约的瘾君子更容易感染艾滋病病毒（Kang, Deren, Andia, Colon & Robles, 2005）。在某些方面，自我中心的分析层次类似于典型的基于属性调查研究，即在受调查者通常的个体特征——例如性别、年龄和教育背景等属性之外，再补充一些可从其直接网络关系中获得的测量值。自我中心网络研究设计比较适合调查那些不太愿意与他人接触的对象。1985 年美国成年人口社会综合调查（Marsden, 1987）率先识别和收集有关受调查对象的客体的信息，我们将在第 3 章详细论述。

第二个分析层次是由成对的行为人构成的对偶网络。假如双方组合顺序无关，就像在婚姻状况中两个人要么是未婚、同居、已婚，要么是分居、离婚等，那么由 N 个行为人组成的样本则有 $(N^2-N)/2$ 种分析对偶组合。但是，如果组合关系的指向有意义，比如下达命令或者征求意见，那么这一样本包含 (N^2-N) 种有序对偶组合。对偶网络最基本的问题是：两行为人之间是否存在具体联系？如果是，那么这种联系的深度、强度和持续时间又是多少？与此密切相关的一个问题是：没有直接联系的双方是否仍然经由中间渠道（比如经纪人、中间人）间接联系？

一般的网络分析都试图将对偶关系的各种变化解释成二元特性的一个函数，比如羽禽群居的同质假设，或者是异性相吸的互补假设。与一般看法相反的是，一项关于1951年至1985年间国际关系冲突的研究发现，具有紧密军事与经济联系的两国和缺乏如此紧密关系的两国相比较，在陷入争端的可能性上前者也只是比后者稍微少那么一点点(Benson, 2004)。

第三个网络分析层次当然就是三元关系了。N个行为人有$\binom{N}{3}$种组合，这是从N个行为人中每次取3个得到的组合数。在3人一组的行为人中，显性和隐性可选择关系的所有可能组合将产生一系列16种独特的三元组类型。在实证网络分析中，一个基本的描述性问题就是在16种类型中观察到的三元组的分布情况，这种综合分布量表也被称为三元关系统计表。三元关系结构的实证研究一般关注情感性连接点(比如喜好、友谊、敌视)，并且特别关注三元关系的平衡和传递(例如，假如A选择B，B选择了C，那么A是否也会倾向于选择C呢)。因为本书没有足够篇幅来评述三元网络分析方法，有兴趣的读者可以关注 James Davis, Paul Holland 和 Samuel Leinhardt(1979)的研究项目，以及 Wasserman 和 Faust(1994:556—602)的全面综述，以获得更加详细的资料。

除了上述三个微观层次的分析之外，全幅网络是最重要的宏观分析层面。研究者利用所有N个行为人的每一种关系的信息，对整个网络的结构关系进行描述和解释。其中，研究者特别关注在此网络系统中行为人的特定位置或者社会角色的展现，以及在这些位置上和位置间行为人联结的模式。虽然一个全幅网络有N个行为人和(N^2-N)种对偶组合(自我关系一般被忽略)，但是所有这些元素加起来只形成单个系统。在全幅网络层面上分析结构变化的原因和结果的各种假设，可能需要利用多个不同系统，而且也是相当耗费研究成本的。然而，全幅网络分析却广受欢迎，其中一些例子包括对爱尔兰都柏林地区青少年朋友关系和伴侣关系的研究(Kirke, 1996)、世界各城市间优势等级分层的研究(Smith & Timberlake, 2002)，以及全球信息产业部门战略联盟的研究(Knoke,待出版)等。

网络分析的这四个层面表明，在某一个层次上涌现的现象，并不能够由其

他层次上关系的现有认识简单地通过逻辑推导出来。例如，选择关系的传递在友谊构成的理论上是一个非常重要的变量（我朋友的朋友，就是我的朋友），这在三元关系层面中可以观察到，但是在自我中心关系和对偶关系层面上就不可能观测到。举另外一个例子：考察两个科学研究团体，他们在科研讨论网络中具有大概相似的自我中心结构、对偶结构和三元结构。但是，假如第一个科研团体的全幅网络分为几个互不相关的小组，那么其中的许多科学家就不可能与其他人接触，甚至间接接触也不可能。假如第二个科研团体的整个网络中各小组间有联系纽带和通道，那我们就可以想象到，其中的信息沟通更加流畅，科学创新的频率更高。网络分析作为一种实证研究的理论指导框架，它涵盖各种突发的结构关系，从而能够解决多个分析层次上的问题。网络分析这一变化多端的功能使得它迅速流行起来。

第 3 章 | 数据收集

对于任何网络实证研究，调查者在开始收集数据之前，必须首先处理三个重要的事项：范围界定、网络抽样、关系测量。由于在网络研究项目里，设定行为人的边界是一个不言而喻的起点性工作，因此我们就从这点开始讲起。

范围界定

关于网络分析的范围界定问题，可以这样简单地设问：在收集社会关系的数据时，也许这些社会关系实际上并不存在一个清晰的界线，那么一个研究者在哪里设定边界呢（Barnes，1979：414）？ Laumann，Marsden 和 Prensky（1983：19）一开始通过对比唯实论与唯名论的定义策略来寻求他们的答案；在设定具体网络行为人的资格以及确认待分析社会关系的类型上，他们制定具体的涵盖规则。但在随后的网络范围界定范式上，他们用三个基本方法取代了上述二分法策略：位置、关系和事件（Laumann，Marsden & Prensky，1989）。本节我们将分析这两种处理过程。

唯实论和唯名论的界定策略　在唯实论的界定策略中，网络分析者采纳系统行为人自身的主观意向，将范围界定在某一实体组织（例如一个家庭，一个公司或者一个社会团体）内的所有或者几乎所有的行为人都可以自觉体验到的限度内。行为人及他们的关系是包含还是排除，得依赖于其他行为人对于相关性的判断。例如，Knoke 和 Laumann（1982：256）即采用这一规则来选择美国能源与健康国家政策领域中的核心组织：“一系列重要的行为人旨在参与起草、建议或选择有关行动方针（即政策选择），以期解决当前明确的重要问题。通过确定

他们之间的相互关联性或者共同目标的实质性定义标准，便得到了一个作为子系统的政策领域。”有些组织在国家能源和健康政策的制定上，被圈内人士认为并不具备什么影响力，因此研究者就将他们排除在外。

在唯名论的界定策略中，为某一特定研究项目的分析或理论目标服务的一个概念框架，在这里被当成是一个先在的理念，用于界定某一网络边界。在许多情形下，法定的或者正式的成员资格要求已经划出了一条明确的界线，比如某一课堂上的小学生、某州律协的律师、在联合国具有普遍协商身份的组织。另一个例子是，马克思主义研究者在定义工人阶级时，将其界定为具有共同生产关系的所有雇员。当然，在何种程度上，主观评价和概念强加这两种界定策略恰好会得到相同的边界，这完全是个实践的问题。

定位策略　这种策略利用行为人的特征、他们在某一正式组织内的成员资格，或者对某一个清晰界定的位置的占有，来对某一网络进行边界划定。Useem(1979)对美国商业精英的研究，便是这一定位策略的范例。他将商业精英界定为在1969年的797家美国最大公司的董事会里的成员，并从中选出了8623个样本。Knoke(2001)在分析全球信息产业部门的公司战略联合时，将公司在1989年至1998年间《财富》500强、《财富》1000强、全球500强的榜单排名作为遴选标准。Feldman-Savelsberg，Ndonko和Yang(2005)调查Bamiléké城6个喀麦隆妇女同乡会成员的集体记忆选择时，要求这些妇女都来自同一个村庄或者部落。研究者利用定位策略可能会发现，某一组织的成员名单或者花名册已经过时、不完整，或不准确。他们需要自己做普查来编制一个完整的成员名单。

定位策略一般会得到一个行为人的集合，他们在某一正式组织中拥有相似的位置，即使有的人彼此之间并没有直接关系。对于使用定位方法揭示的网络结构的代表性，研究者应该保持足够的警惕。因此，商业精英间的联系与相对低阶位的雇员间的联系是大大不同的。跨国公司间形成的战略联盟与小企业之间的合伙也是大不相同的。伴随定位策略而生的另一个网络问题是，行为人之间往往并无联系，或者构成一些小的紧密团体但彼此间缺乏联系。例如，有些喀麦隆妇女，虽属同一同乡会，但是通常彼此间却完全陌生(Feldman-Savelsberg, Ndonko &Yang, 2005)。

在采用定位方法进行网络分析时,研究者应当对涵盖或排除某一特定定位标准提出明确合理的理由(Laumann et al.,1983; Scott,1991:58)。研究者可以采用唯名论标准,为他们的排除规则设定一个主观性门槛,即使某一定位处于持续变化之中。例如,为了研究大公司,研究者也许需要从《财富》榜单中抽取 50 家、100 家甚或 500 家公司。如何设定选择范围,更多地取决于时间和研究经费限制,而不是取决于入围公司与未入围公司的某些实质性差别。Galaskiewicz(1979)在对一个叫 Towertown 小城的一些组织进行筛选时,就生动地说明了上述过程。首先,他采用属地原则,即将研究限定在一个地理范围内的人口。其次,他根据他的研究时间和经费限度,利用行业标准,排除了商业机构、交通业、公用事业、地产业、街道居民组织、社区机构以及小学等。

关系策略　这一方法依赖于知情者或网络行为人本人提名添加其他行为人。关系策略包含几个常用的做法,比如声望法、滚雪球抽样法、固定名单选择、扩展性选择以及 K 核心方法。我们将在本小节简要地描述一下这些关系策略方法所需的条件及其局限性。

在声望法中,研究者通常咨询有关资深知情人士或者专家,请他们提名有关行为人。例如,调查者调查两个精神健康治疗机构网络,首先统计当地社区精神健康治疗机构总名单(Morrissey, Tausig & Lindsey, 1985)。接着,他们会询问调查对象,请他们从总名单中挑出 5 个他们认为最重要的,或者他们最经常联系的机构。通过这两个步骤,便挑选出了两例核心的精神健康治疗机构。接下来,再请关键知情者即人道服务计划委员会和矫正部门的主任们,从总名单中再挑选额外的重要机构,对核心名单进行补充。

声望法非常依赖关键知情人提供准确全面的信息,因此研究者在确定重要知情人士上的能力便非常值得关注(Scott,1991:59)。Morrissey 等人(1985:35)就曾经忠告:“在评估声望法划分标准的准确性上,并没有一个统一的准则。很显然,不同的划分标准会产生不同的行为人涵盖范围,也会导致随后的分析受到影响。”对采用声望法获得的网络关系名单的可靠性和完整性进行评估,是一件非常困难的事情。经常只有等到一个项目完成之后,才可能对此进行评价。因此,网络研究者挑选关键知情人,应当有充分的理论和实践依据,且这些

理由是独立于受调查的具体社会关系的(Scott，1991：59)。但有时抽样偏差问题——即由于非随机性的案例缺失而导致样本不能代表目标总体(Allison，2001：78—81)——会超越关键知情人士的遴选问题。关键知情人士不管如何博学多才，他们所提供的数据，总是与通过滚雪球抽样法所得到的数据有系统性差别。

在滚雪球抽样法中，研究者从小范围的网络行为人开始，由他们来提名与其有某种特定关系的其他参与者。同样，这些其他参与者也被要求提名其他行为人，直到没有或者几乎没有其他人名再被提及(Frank，2005；Wasserman & Fraust，1994：34)。在此方法的早期严格版本中，每一轮滚雪球抽样都将产生等量的行为人(例如，列出3个你最好的朋友)，并且在每个阶段每一个行为人都被问及同样的问题(Goodman，1961)。最近，这两个特殊条件有所松动，允许使用不同的问卷，提出数量不等的行为人(Wasserman & Faust，1994：34)。因为滚雪球抽样法利用网络行为人的社会关系来构建样本，所以每一轮都会提出新的行为人，而这些新的行为人与现有的行为人存在特定关系。因此，滚雪球抽样通常会得到一个关系非常紧密的社会网络(Laumann et al.，1983)，因而被称为连锁推荐法(Heckathorn，1997)。应答驱动抽样是该方法的一个复杂延伸，它将滚雪球抽样法与某一数学模型结合，从而增加抽样的代表性，以弥补非随机性选择所带来的偏差(Salganik & Hckathorn，2004)。

在社会网络分析中，滚雪球抽样法对寻找难以接触人群具有独特效果，比如毒品交易者、吸毒者、非法移民、HIV阳性者的性伴侣、地下抵抗组织等。由于缺乏对此类隐蔽人群的抽样框，常规抽样调查方法经常得不到足够的样本量。相反，滚雪球抽样法可以从访谈一个初始样本开始，比如瘾君子经常光顾的治疗机构或者街区。受访者被要求提供与之有特别关系的其他行为人的地址信息。接着，调查者联系这些被提名者，并且请他们提供与其相关的其他参与者。因为网络数据收集工作常常需要知道行为主体介绍的行为客体的身份，所以获得其知情同意、保护匿名来源、确保隐私等要求就引发了一些伦理问题(Borgatti & Molina，2005；Klovdahl，2005)。如果研究者直接利用滚雪球抽样法来招募行为客体，人权事务委员会也许就不会批准这种抽样设计方案。作

为替代办法，应答驱动抽样设计会要求种子行为主体告知他们的行为客体，如果自愿参与项目可以得到一点小小的补偿。

在实施网络抽样时，一个程序性问题——是否给知情者提供一个人员名册供其选择还是让受访人自己提名——将选择方法分成了两种：固定名单选择法和扩展选择法(Doreian & Woodard, 1992)。在固定名单选择法中，受访人被限定于在研究者和知情人士事先所确定的行为人名单中，报告其社会联系。在扩展选择法中，受访人可以不用参照某一名单而是按照自己的意愿确认其他行为人。扩展选择法的运用非常类似于滚雪球抽样法的程序。在 Doreian 和 Woodard(1992)的一项关于小孩和未成年人服务体系项目机构的研究中，任何得以涵盖在此网络关系中的服务机构必须满足下面两个条件中的一个：现任机构主管至少引用 3 次，或服务员工至少提名 5 次。低于上述提名门槛次数的机构将被淘汰出局。Doreian 和 Woodard 的报告中指出，固定名单选择和扩展性选择得到的网络结果，在几个方面都存在很大差异，其中包括行为人数量、关系节点数量、关系点的密度和质量等(关系点的密度是通过将观测到的双向关系点数除以所有可能的关系点数得出来的，关系点的质量是知情人士对关系的满意度、可延展性及价值的估量)。特别需要指出的是，固定名单法只从总行为人中选出 50%的行为人，40%的双向关系节点是通过扩展性选择方式发现的。固定名单选择方法更倾向于产生非随机性抽样偏差，也就是说，它总是产生一套核心行为人名单，而系统性地排除了边缘关系行为人。这个方法会有个不好的结果，即得到一个没有结构性背景的关系网络。除非条件可以确保两种方法都得到一个相差无几的网络，否则固定名单选择法并不能作为扩展选择法的替代方案，即便它有管理成本低的优势。

最后一种关系策略是 K 核心法。Seidman(1983)提出了 K 核心的概念，用于发现在一个大网络中的行为人子集。这些行为人子集一般包括许多的小团体，小团体之间关系微弱，但是在团体内却团结紧密(Yang & Hexmoor, 2004)。假如在一个小团体中，每一个行为人与至少 K 个其他行为人存在联系，则这个小集体就是一个 K 核心。通过改变 K 值，研究者可以为某一网络的边界设定或宽或窄的限制标准。Doreian 和 Woodard(1994)运用扩展选择法，描述

了K核心概念是如何用于定义和确定网络边界的(我们将在本书第4章更详细地讨论K核心法)。通过改变K的阈值,研究者能够使一个非常大而松散的网络边界重新得到划定。当K值高时,边界标准更严格;当K值低时,边界标准更宽松。为了实用目的,Doreian和Woodard(1994)推荐使用一个低K值,以便建立一个更全面的网络边界。相对来说,一个涵盖范围更广的网络更少受到抽样偏差方面的质疑,同时研究者随后总是可以运用高K值来构建一个更加严格紧密的关系网络。然而,一个涵盖范围大的网络却会带来更重的数据收集任务:虽然低K值会使得网络范围更宽,但是研究者在每一轮都必须访谈数量巨大的研究对象,而这是一项花费巨大且容易犯错的任务。降低K值门槛也会给整个网络节点的数量带来指数增长而非线性增长。虽然目前K核心法的实际运用非常少,但是我们相信,在界定网络边界的实证研究中,它具有巨大的潜在价值。最近网络分析软件的发展应该便利了K核心法的使用。

基于事件的策略 这一方法在划定网络边界时,根据发生在特定事件和地点中的行为参与者而定(Marsden, 2005; Scott, 1991:60)。例如,Freeman和Webster(1994)在南加州海滩观察休闲人士时,将范围限定在一个月内至少有3天会来海滩的人士。他们共收集了353个事件的参与人,其中包括43位规律性光顾海滩的爱好者。基于事件参与的涵盖规则完全不同于位置策略法和关系策略法,后两者都忽略了行为人的行为。

由于基于事件的策略依赖具体的行为或事件来确定网络边界,因此如何对这类行为进行合理归类就显得非常关键。对于基于事件的分析来说,最让人畏惧的便是为选择能够回答具体研究问题的事件提供充分的理由。研究人员能够涵盖那些中立观察者认为是重要的,或者知情参与者认为是重要的事件。基于事件的边界界定方法易于因忽略某些重要的行为和参与人而导致获得不完整数据或者缺失数据。对于依靠一个单独事件来界定网络边界的研究者来说,这个问题尤其敏感,因为许多重要的行为人也许根本就没有参与该事件。因此,对多个事件进行观察,一般来说也许可以获得一个更加全面的分析网络。由于每一个事件都包含着一个潜在的独特的网络小团体,因此同时考察多个事件便得到一个可延展的网络边界:每一个事件都产生一个独立的网络,而其中

的参与者部分地与其他事件的参与者相重叠。将所有事件的参与者集合起来，便得到一个更加广泛的网络，这一网络能够更好地回应研究者的问题。例如，为研究大学宿舍里的学生互动情况，Freeman 和 Webster(1994)对咖啡馆和宿舍这两个不同环境中的社交聚会的事件参与者进行了观察。

数据收集过程

这一节讨论构建数据库的几个基本的网络数据收集过程，包括单个和多个提名生成法，全部个人网络关系的测量，位置选择法和资源选择法，以及档案资料的使用。

单个和多个提名生成法　提名生成法常用于自我中心网络关系研究，其通过调查问卷从每一个主体受访者那里收集与主体有直接联系的关系客体间的相关关系的信息(Marsdenm, 1987)。自我中心网络关系研究者依赖于两种调查工具：提名生成法和姓名阐释法。提名生成法确定调查对象的关系客体，姓名阐释法则获得每一个关系客体自身、关系客体之间的关系，以及其与关系主体之间的关系的信息(Marsdenm, 2005)。首先，主体受访者被要求提供一定数量的人名，这些人名与其有某种特别联系。接着受访者会被问及他/她所提供的人员之间是否有特别联系。最后，受访者还得提供每一个关系客体的具体特征，比如年龄、性别、种族和教育程度等。

根据研究目标的不同，自我中心网络研究可选择利用单个或多个提名生成法。单个提名生成法依靠一个调查问题来引出关系客体的名字。1985 年美国综合社会调查中关于美国人的核心讨论圈的问题模块，生动地体现了这一方法设计(Burt, 1985; Marsden, 1987)。在 GSS 问卷中，为了引出受访者的知心朋友网络，访谈员会问：

> 大多数人总是会与其他人讨论一些重要的事情。那么，请你回顾一下过去 6 个月当中，你与谁讨论过重要的事情？你只需要说出他们的名字或者名字缩写就可以。

访谈员记录每个受访者说出的6个名字，然后继续问到："你感觉跟所有这些人的关系都是一样的吗？（如果不）那么其中哪一个是你觉得关系特别近的？（可以接着试探再问：还有其他人么？）"之后，他们会问受访者关系客体之间的联系：

> 请想想你刚才提到的那些人之间的关系。一些人之间也许完全陌生，他们即使在大街上碰在一起，互相也认不出来。一些人之间关系也许特别亲密，跟他/她和你的关系一样好甚至更好。那么，现在想想甲和乙。他们之间完全陌生吗？或者他们关系特别亲密么？（试探问：跟他们和你的关系一样亲密？甚或更亲密？）

受访者还要求回答：他们与关系客体认识多长时间了？平均间隔多长时间交谈一次？与受访者之间是什么关系（比如配偶、父母、子女、邻居、同事、朋友、顾问）？同时他们也要报告每一个客体的性别、种族、教育背景、年龄、宗教和政党认同。

1985年GSS的受访者平均只有2.94个可与之讨论重要事情的人，而这些核心讨论圈成员的55%是家庭亲属（Marsden, 1987）。自我中心网络密度测量均值为0.61，意味着多数关系客体相互之间认识（自我中心网络密度测量关系主体所感知到的关系客体间的关系强度，其中0值等于完全陌生，1值等于非常紧密，0.5值等于其他）。美国人的核心讨论圈在年龄和教育背景方面比一般人群更具有同质性。性别多样性均值为0.68，意味着在绝大多数核心讨论圈中男女性别是比较平衡的。相反的是，种族/民族异质性均值为0.05，这一低值意味着绝大多数关系客体都是同一种族的，Marsden（1987）将这一现象归因于大部分的亲属间提名。

2004年美国综合社会普查沿用了1985年GSS的提名生成法，但是网络规模均值从1985年的2.94下降到2004年的2.08，这是一个主要的变化（Mcpherson, Smith-Lovin & Brashears, 2006）。1985年，一个受访者平均有3个知心密友，但20年之后却变成了0：几乎近1/4的受访者无人可与之交流重

要事情。非亲属关系连接点的巨大下降导致社会网络更加集中在配偶与父母中，而与公益组织和邻居的联系越来越少。受访者与关系客体的同质性持续偏高，伴随着教育背景异质性的下降，种族异质性上升。分析推测，美国人口特征的变化某种程度上可以解释美国人与日俱增的社会孤独感。

Ruan, Freeman, Dai, Pan 和 Zhang(1997)利用 GSS 中的单个提名生成法于 1986 年和 1993 年在中国天津进行了两次调查。在这两次调查中，天津受访者普遍提出了更多的核心讨论圈成员（均值分别为 4.58 和 3.30），这比美国 1985 年的均值 2.94 高出许多，而且天津人的核心讨论圈成员中亲属要更少一些。相对于 1986 年，1993 年的天津受访者核心讨论圈中较少提及同事和亲属，但是更多提到朋友。这些变化反映了中国自 1978 年以来社会宏观结构的变迁，其间中国利用更加灵活的市场雇佣制不断替代终身雇佣制。伴随着变迁的持续，人们逐渐发展了更多工作圈之外的社会关系，因此导致核心交往圈中其他关系人的加入机会的增加(Ruan et al., 1997)。

1985 年 GSS 调查问卷在利用提名生成法时并没有关于所谓“讨论重要事情”的内容提示，而是将解释这一术语的任务交给了受访者。这种模糊处理可能使不同受访者对此有不同解释，从而影响自我中心网络的结构或组成(Bailey & Marsden, 1999)。例如，受访者如果将“重要事情”界定为个人和家庭事务，那么他将更多地提名联系紧密、有亲属关系的网络成员，而将“重要事情”界定为与工作或政治有关事项的受访者则较少提名亲属成员。为了仔细研究受访者在解释“讨论重要事情”时的认知过程，Bailey 和 Marsden(1999)使用共时有声思维方法，对 50 个受访人进行调查。首先问这些受访者提名生成法使用的标准问题，接着立即用各种试探方法研究他们的思维过程。大多数受访者将之解释成个人问题，如家庭问题或个人间的问题。虽然先前的调查问题会明显引导受访者用其他方式解释“重要问题”，但是这些不同的解释并不会形成实质性不同的网络构成。为了方便将来运用单个提名生成法，Bailey 和 Marsden(1999)提出了 4 种可选策略。第一种策略将定义内容与诱导关系客体分开：受访者首先被要求根据自己的理解定义重要事情，然后根据其定义来提名客体。第二种策略使用举例法来解释重要事情：研究人员提供一些例子，来帮助受访者定义有关

重要事情。第三种策略中，研究人员为每一位受访者明确解释重要事情的含义。第四种策略则利用调查问卷问题次序的不同排列，来弱化有关问题背景对受访当事人理解重要事项的影响。

虽然单个提名生成法产生一个关系客体的核心集合，但是它所得到的联系人只是利用多个提名生成法所得名单的一部分。研究人员通常会对单个提名生成法所得的核心问题之外的各种常规行为感兴趣。在一个关于"朋友"概念的有趣研究中，Fischer(1982)发现人们使用这一相同的概念来描绘完全不同的关系。Fischer 没有采用单个提名生成法，而是采用了多个提名生成法。这一方法包含 9 个事项，如家务、询问合适的贷款数额、社交、讨论有关工作、业余爱好以及其他个人事情等等。在北加州的调查中，他从每一个受访者处获得数量不等的客体名单(从 2 个到 65 个)，平均下来每个受访者提供了 18.5 个名字。而且，受访者将其中 11 个人视为朋友，很显然各人对于"朋友"的定义五花八门，包含了各种人际关系。在以色列进行的同样两项调查，平均得到 14 个和 11 个姓名(Fischer & Shavit, 1995)。结果发现，以色列人比北加州人有着更高密度的自我中心社会网络。

为了比较单个和多个提名生成法，Ruan (1998)在中国天津调查时使用了 11 个提名生成法。除了利用 GSS 中的重要问题项外，她还利用了一些工具性连接点，比如家务和借贷，还有一些表达性关系，比如社交、谈心和寻求建议。GSS 提名法平均得到 3.30 个天津人的关系客体姓名，而其余十种提名法总共得到 8.17 个名字。而且，受访的中国人将"讨论重要事情"当成是社交性表达事项，大多数受访者将与其社交的朋友(出去聚餐、购物或者游玩)视为他们讨论圈中的成员。相反，由工具性提名生成法所得到的名单更少被涵盖在核心讨论圈网络中。

由"认识"提名生成法可得到最大的自我中心网络。这种提名法要求受访者报告其认识的所有对象。Killworth, Johnsen, Bernard, Shelley 和McCarty (1990)的研究表明，在美国和墨西哥，这种认识提名法可以产生 1500 名认识对象。然而，为了获得与受访者有实质性联系人员的可操作名单，必须对此施加一些严格条件。Campbell 和 Lee(1991)总结了 4 种限制提名生成法的典型条

件:(1) 角色/内容限制,将受访者限定在一种或几种关系中进行提名;(2) 地域限制,将受访者所提关系人限定为居住于一定区域;(3) 时间限制,要求受访者所提关系名单限定在一段可回溯的时间段内;(4) 数量限制,让受访者在符合标准的人员名单中只提出 N 个关系人(比如,3 个你最好的朋友)。大多数提名生成法综合采用上述 4 种限制条件的其中几种。例如,Campbell 和 Lee(1991)请 690 位纳什维尔受访者利用其所居住位置地图,说出离他/她所知道的最近的 9 所或 10 所房子邻居的姓名。之后他们又要求受访者确认在过去的 6 个月中与之聊过至少 10 分钟或到访过该家的邻居。相比其他几种提名生成法,1985 年 GSS“讨论重要事情”提名生成法在内容、数量和时间段上的限制最为严格,因此也只得到一个网络密度最高的最小自我中心结构网络规模(均值 3.01)。Fischer的北加州调查,利用多个提名生成法,获得了一个最大规模的网络(均值为 18.5),但是其网络密度也最小。纳什维尔调查中的关系持续时间是最长的,因为没有设置空间方面的限制,GSS 调查中的关系联系频率是最高的。Campbell和 Lee(1991)的研究使人们了解到,通过提名生成法所施加的限制条件是如何塑造一个目标网络的结构。

位置选择法和资源选择法　社会资源理论的基本假设是,社会结构——由财富、权力和地位来定义——是金字塔型和分等级的,社会资源及其取得路径正是嵌在这种社会结构之中(Lin, 1982; Lin & Dumin, 1986)。虽然有许多研究考察了社会关系连接点是如何成功地催化功利性行为的,但是一直以来却很少有研究者对取得社会资源的路径进行分析。为了解决如何衡量行为人的社会资源这一实证问题,位置选择法要求受访者报告其是否在某一社会位置上与他人具有联系。处于社会等级结构中的某个位置是社会资源的一种合理指标,在此意义上调查某个位置上的人的个人交往关系不仅可以显示其拥有何种社会资源,还可以显示其获得社会资源的途径。

如何通过位置选择法反映社会资源获取途径,主要依赖于社会位置的选择。Lin 和 Dumin(1986)从 1970 年美国普查职业分类指标中选择了 20 种频次最高的职业。这些职业在白领和蓝领工作领域的上层和下层均有所分布。受访者报告其是否在这些职业位置上有任何社会关系,比如亲戚和朋友(指有较

紧密联系的人)或者认识的人(指有较弱关系的人)。利用1975年数据库(含位置选择法)来考察寻找工作的过程，Lin和Dumin(1986)发现关系力度与较好职业的获得这二者之间的联系，与受访者的出身是相关的，比如受到父亲职业的影响。当受访者出身社会上层时，强的和弱的社会关系连接点对于获得较好职业的作用基本平等。但是当受访者出身地位低时，弱社会连接点比强社会连接点能提供更高社会地位的工作机会。随后，Lin，Fu和Hsung(2001)将位置选择法用于研究台湾的职业声望与收入的关系。他们的报告指出，台湾社会在获取社会资本方面呈现出性别不平等现象。社会资本更有利于男性获得声誉好的职业和高的收入，然而对于女性来说，她们更依赖于人力资本(比如教育)来获取较好的工作和收入。

哪些职业要涵盖在研究范围之中，这是利用位置选择法进行研究的关键点，当然这得根据研究问题来定。不同社会阶层的文化差异与社会网络关系是如何联系的？Erickson(1996)利用位置选择法对这一问题进行了研究。她抽取了三个主要的社会阶层的维度，即对财产、组织和技能的控制。由于她的受访者全部都在安全行业工作，因此基于技能控制的19种职业均反映了他们安全行业的特性(8个专业工人，4个蓝领工人，4个警察，2个商业管理者，1个企业主)。访谈员给受访者一些提示："现在我要问你，你是否认识多伦多地区某一行业里的人？比如，你是否认识任何律师？请列举一些你知道能与他们聊得来的人员，即使你与他们并不亲近。"假如受访者回答说他们认识某个职业里的人，访谈人就会问他们关系的亲密度，比如是否"仅仅认识"，是"作为朋友关系的那种认识"还是"作为亲属关系的那种认识"。在一项基于性别的社会资本的研究中，Erickson(2004)根据加拿大普查的数据，特地挑选了15种男性主导或女性主导的职业。她问受访人是否认识每种职业中的女性或男性。Erickson(2004)发现，男性认识由女性主导职业中的人的可能性比女性认识由男性主导职业的人的可能性更大。而且，因为男性在许多社会部门都占据较为重要的职位，所以男性在社会网络中的优势难以改变，尽管相对于女性，两种性别都更倾向于与男性有更加多样的关系联结点。

位置选择法得到的自我中心结构网络，能用于测量个人在社会等级中与多

个职业的连接关系。社会资源依照职业结构的金字塔型和分等级的不同层级进行分配，在此意义上位置选择法所得到的不仅仅是人们获得不同职业的途径，还是人们获得不同社会资源的途径。然而，人们常常获得超层级的关系客体提供的功利性和表达性帮助，而位置选择法却将选择限制在某些职业之中。研究人员常常对个人的社会资本宽泛地定义为人们从各种社会联系中获取的所有类型的帮助。资源选择法主要关注个人社会资本，由函数表达为 $SC = \sum_j S_j$，其中 j 指社会资源项，S_j 指测算这类资源的可获得性(Van der Gaag & Snijders, 2004)。与位置选择法询问当事人是否与特定职业有联系不同，资源选择法询问人们是否认识可利用某种具体资源的人。Van der Gaag 和 Snijders (2004)询问荷兰的受访者，"请问您是否认识谁可以帮助做……"(有 35 种选项，从"修自行车"到"礼节性拜访")他们报告说，17 个资源选择项目形成了 4 个亚层次，这些亚层次具有内在的同质性，但可关联性较弱。其中测量获得教区牧师这样有声望的职位的亚层次，与 Lin 和 Dumin(1986)的位置选择法非常相关。但是，获取信息和获得工具性支持比如帮助搬家)的 2 个亚层次与位置选择法相关性很弱，因此形成了独特的社会资本维度。

测量全部个人网络　一些研究人员对测量全部个人网络感兴趣，因而创造出多种提名生成法来达到目的。全部个人网络指的是行为主体认识的所有关系。依据具体方法，全部个人网络规模可达 250 个至5000个(Freeman & Thompson, 1989; Killworth, Bernard & McCarty, 1984)。

在名单核对法中，研究人员首先随机选出几个姓名(或姓或名)，然后将这些姓名读给随机挑选出来的受访者，问他们:"你认识叫这个名字的人么?"当受访者说认识某一个名字时，访谈者便问他/她有关该人名的信息。McCarty, Bernard, Killworth, Shelley 和 Johnson(1997)利用一组 50 个名字的名册访谈了 793 位佛罗里达民众，采用的就是这一方法设计。访谈者请受访者报告与名册里名字相同的朋友的联系。研究要求受访者通过外貌或名字知道对方，并且他们在最近两年里有过联系。McCarty 等人(1997)通过这种办法获得了最多 14 个关系客体的一个名单，但发现这个名单样本中黑人、西班牙裔和亚裔的代

表性偏低。然而,他们认为在发掘一个更具代表性的全部个人关系网络样本上,这一方法依然优于其他办法。认识名单所列某一名字的受访者比例,与美国人中拥有同一名字的人口比例,二者高度相关,这正表明该方法的偏差是不严重的(McCarty et al.,1997:311)。

在反向小世界方法中,研究者首先创造一个虚拟的目标,其中包括名字拟造的人员、随机分派的地址、年龄、性别、爱好、成员资格、职业以及其他社会经济特征。之后他们要求知情人提供有关的关系客体,而这些人员能直接提供一个信息给虚拟目标,或者形成一个指向目标的网络连接点。明确了关系客体之后,研究者便向知情人士也就是受访者询问这些关系客体的个体特征、与知情人的关系、关系客体之间的认识程度。Killworth 等人(1984)利用 RSW 方法访谈了佛罗里达杰克森维尔的 15 个知情人士,获得了 100 个虚拟美国人和 400 个虚拟国际人士的信息。这些知情人士提供了大约 134 个关系客体,通过这些客体与这 500 个虚拟对象联系起来。大多数关系客体(86%)是知情人士的朋友,而且超过一半都是男性。

Freeman 和 Thompson(1989)认为,RSW 方法只能获得一个人全部关系网络的一部分,因为知情人不可能提出超越虚拟目标数的其他关系客体。为了解决这个问题,Killworth 等人(1990)将 RSW、名单核对法和 GSS 提名生成法综合起来。他们发现,佛罗里达人的个人全部网络关系规模平均是 1700 人,而墨西哥城居民的平均值是 600 人。利用同样的数据资源,Bernard 等人(1990)调查了上述不同方法中的重合部分。他们发现 GSS 提名生成法和多重社会资源提名法总共才获得 RSW 提名法所得个人全部网络关系数量的 18%。名单核对法(通过姓氏核对)得到的个人网络关系数量最大。

档案资料　档案资料数据为社会网络分析提供了许多有用的信息,虽然这些信息原先并不是为了网络研究而收集的。与专门调查相比较,档案资料相对而言使用成本低,对受访当事人不会造成时间和精力上的负担,而且当数据是重复收集的话,档案资料有可能保留一些非常难得的纵向信息。档案资料来源形式多样,包括网页、个人信件、专利摘要、书籍和论文文献,以及电脑网络交流文本等。几个在线数据供应商,比如 Lexis/Nexis, Free Edgar, SDC Platium

以及美国专利商标局等，存储了大量有关公司董事会、战略合作伙伴、小企业主以及专利摘报等信息，极大地方便了研究人员对商业组织网络的研究。

档案资料检索新技术的出现总是层出不穷。利用因特网这一巨大丰富的个人信息库，Adamic 和 Adar(2003)设计了一种技术，用来发掘其中的人际关系网络。他们用一种客户端软件，调查两个大学(斯坦福和麻省理工)服务器上个人网站的网页内容。他们关注每一个主页的文本、外部链接、内部链接和邮件名单。他们这种发掘因特网内容的创新尝试，不但了解到谁认识网站主人这类信息，而且还了解到他们的认识是在何种社会情境中发生的。例如，麻省理工的学生倾向于通过大学生联谊会及其下属组织来互相认识，其中提名最多的 10 个社交场合有 5 个是大学生联系会之类的组织。比较而言，斯坦福的学生之间的认识渠道更加多样化，比如通过研究团体、宗教组织、文化交流机构以及学生联谊会等。

个人书信和电子邮件带有大量有关个人交往的信息，因此分析检查这些资讯文本会得到有关个人网络和社会结构方面的丰富信息。这种方法对历史学家尤其有利，因为他们所研究的古代社会现在是不可能去调查的。例如，Alexander 和 Danowski(1990)通过 Cicero 的书信，对古罗马的社会结构进行研究。Cicero 是古罗马著名的演说家和有影响力的政治家，他跨越两大社会精英阶层：无官职的“骑士阶层”和有官职的“元老院阶层”。研究者查阅了 280 封 Cicero 18 年间与他的熟人、朋友、亲属的通信。他们的数据管理程序记录了 524 个人之间的 1914 种关系，包括关系双方的姓名和官职，以及诸如送礼、预约、中间调解、帮助、负面描写、拜访等关系内容。该研究一个主要的发现就是，虽然元老院成员和骑士在某些事情上互相对立，但是他们似乎构成了一个单一的、整体融合又互相牵制的社会阶层。

期刊和专利都含有大量引文名单，因而为社会网络研究提供了另一个很好的数据来源。Hargens(2000)对 3 门自然学科、3 门社会学科、1 门人文学科的引文模式进行了研究。对每一门学科，Hargens 首先确定一个期刊清单，这一系列期刊使用引文索引服务和专家咨询服务。然后他记录每一篇论文中引述其他文章的参考文献或脚注，从而为每一个学科领域建立了一个全面的引用关系

网络。Hargens 发现研究者因各种目的而引用他人文献，从认同一个普遍的理论流派到纠正某一具体的计算过程。特别需要指出的是，自然科学的文献引用大多数是为了拒绝或更正某一具体实验方法，而社会科学家绝大多数是为了认同或挑战某一学科的基本假定。在对之前的文献进行争论方面，人文批评领域往往要比自然科学和社会科学界更为盛行。

对于有志于研究知识流动的人来说，专利授予过程产生了可资研究的丰富数据。在对专业技术市场的研究中，Podolny 和 Stuart(1995)深入分析了专利摘要的文献引用。为了获得专利，技术发明者要向美国专利局提交申请书。作为申请程序的一部分，他们必须列举出作为此次申请重要基础的所有已授权专利。专利审查员需要确认引文清单包括了所有之前的相关发明。利用 Lexis/Nexis 在线数据库，Podolny 和 Stuart(1995)分析了自 1976 年至 1991 年间授予全球半导体技术创新者和制造者的所有美国半导体设计专利。他们将专业技术市场看做是由一系列技术创新构成的，这些技术创新之间是由一种创新到另一种创新的知识流动链进行连接的。因此，每一项授权专利都是一个进入已有技术市场的新的重要的技术创新。一项创新技术的专利摘要文献，提供了与之相关的所有前人的基础性创新信息，而基于此项发明的后一项专利也表明它们之间的联系是建立在此项重要创新之上的。因此，一个专业技术生态位构成了一个自我中心网络，该网络是由某个重要创新及它与其他专利发明之间的联系形成的。

虽然调查文献引用清单可以构筑学者或专利人之间的关系网络，但是其中有两点重大差异(Meyer, 2000)。第一，期刊论文倾向于引用大范围的已发表文献，而专利引用范围较小，往往只引用对此项创新有重要贡献的前人发明。第二，期刊论文引述目的较广，比如赞赏有关论著、纠正前人工作或者与某人争论某个观点(Meyer, 2000)。相反，专利摘要文献引述的目的就单纯多了，就是为了感谢前人的工作为当前的创新提供基础。因此，利用专利文献和期刊论文所建构的社会网络可能在内容上有根本性差别，需要研究者仔细甄别。

政府机构和经济组织在雇员间的交往以及有影响力的组织间的互动上，已经累积了大量的信息。然而，这些有价值的数据资源普遍被忽视，组织理论者

对此也少有问津(Perrow, 1986:172—173)。一个例外是 Baker 和 Faulkner 的研究,他们利用了 Kefauver 特别委员会报告。这一报告是该委员会在美国参议院反垄断司法委员会上关于商业合谋的证词。通过这个报告,作者揭示了非法商业关系网络的结构及其动因。

自 20 世纪 90 年代以来,随着电脑网络技术的迅速应用,电子化交流不再局限于少数技术专家,而是迅即扩展到普通民众。新技术用于控制或激励网络使用者的交流,从而让研究人员便于研究它、揭示它的模式。社会网络研究者通过检视这些电子化交流(CMC)系统,比如电子邮件、电子公告栏、语音信息和其他交友软件,获得了丰富的社会网络数据(Rice, 1990, 1994; Rice, Borgman, Bednarski & Hart, 1989)。由于监控程序普遍都可存储所有的 CMC 系统的交流,因此它的网络数据便构成了此系统使用者间所有交往的一个普查。然而,对于网络分析人员来说,检查分析每一项交往,在计算时间和成本上将会是不可承受的。更为重要的是,由于这种电子化交流在内容、时间和目的上都呈现多种形式,因此并不是所有的交流数据都是有用和相关的。CMC 研究者通过 3 个原则甄别数据:(1) 在基于角色的研究中,研究者过滤那些基于角色——比如经理或者会计——的信息发送者或者接收者;(2) 在基于时间的研究中,研究者设定一个时间段,在此时间段内收集所有交流信息;(3) 在基于内容的研究中,研究者过滤一些内容和主题,而只选择那些符合研究问题的信息。客户端计算机软件通过关键词对大型网络数据库进行快速搜索,极大方便了基于内容的研究。

与自填式问卷调查相比,CMC 分析方法的强制性较弱,且具有更小的测量误差。例如,当两个人自填的关系不一致时,要发现这些测算错误的来源,比如回忆不准确、回答有偏、问卷用词问题等,就变得几乎不可能了(Rice, 1994: 176)。然而,计算机控制网络数据能够精确地反映 CMC 信息的内容,因此有关相互关系的假定就能够得到明确的区分和验证。虽然 CMC 数据比自填式问卷得来的数据更精确,但他们并不一定更有效。他们通常展现人类交流的不同方面(Rice et al., 1989),例如,电子邮件交流能更准确反映工作行为而非朋友间和家庭成员间的社交行为。作为使用工作服务器电子邮件系统协议的一部分,

雇员普遍都签署了协议同意自愿分享他们的交流。因此,依据研究主题的不同,研究者应当对CMC系统产生的数据保持警醒。对某一公司服务器用户的电子邮件交流进行监控检查,可以获得关于工作联系结构的丰富数据,但是却很难获得雇员业余时间的关系结构数据。单独依赖CMC系统也可能会错失一些非常重要的细微交流之处。大量信息内容是不重要的例行公事,偶尔会有一条非常重要的信息,忽略这样的信息也许会产生严重后果。因此,研究者应当对CMC数据分析保持警惕,因为这种方法虽然揭示了有价值的背景线索,但同时也丢掉了一些交流行为中的重要的东西(Rice, 1994:177)。

CMC技术的最新发展大大延伸了网络的运用,便利了社会网络分析(Berman & Brady, 2005)。网络基础设施包括诸多网络工具:比如高性能计算机和网络,个人电子设备(即时信息工具、黑莓手机和个人数字化助手),数据获取、管理和分析服务,可视化设施等。网络基础设施为社会网络分析提供了大量新的机会。例如,高性能计算机服务器记录万维网和个人电子设备使用者的文本信息。对这些文本信息进行分析会得到重要的社会交往模式和网络构造。大量的机会同时也带来巨大的挑战,从如何获取、存储、分析海量数据这些技术问题,到如何确保获得知情同意权以及保护隐私等法律和社会的问题。因此,网络基础设施所支持的社会网络分析,要求多学科之间进行前所未有的合作,包括社会科学、法学、电子学和计算机科学等。

Adamic和Adar(2003)利用网络工具获取和分析巨量数据的研究就是一个很好的范例。这些数据信息在网络空间流转,通过对它们的分析,深化了我们对社会网络中某些重要事件的理解。为了研究小世界现象,他们从惠普公司实验室员工的电子邮件和超过2000名斯坦福大学学生的社交网页中,收集了大量信息。这个交流网络包含流通在430个人之间的4000封电子邮件。这些人在3个月的观察期内,每人至少交换过6封电子邮件。网页网络数据来源于一个社群网址Club Nexus,它邀请斯坦福大学学生加入该俱乐部并且登记他们的朋友圈。登记过程中也要求登记人报告他们的专业、性别、个人特征、宿舍地址、爱好等。朋友也被请求回应登记在Club Nexus的关系。未得到回应的关系将从数据库中删除。为了定位交流目标,Adamic和Adar对被观察电

子邮件网络使用了3种搜索策略，每一种都反映了一个潜在的理论观点。高相关度寻找法选择那些通过认识很多人来了解研究对象的个人。第二种便是将信息传递给在组织等级架构层级中离研究对象最近的联系人。第三种便是将信息传递给在空间距离上离研究对象最近的联系人。在两种网络中模拟这三种策略后，Adamic 和 Adar 认为第二种——位置搜索发现了实验室网络中关系双方的最短路径。然而，没有任何一种搜索方法得到了俱乐部网络的最短的合理路径。

认知社会结构

虽然关于社会网络最普通的问题之一可能是："谁认识谁?"但是，与认知社会结构(CSS)分析最相关的问题也许是："谁知道谁认识谁?"研究者使用 CSS 数据调查知情人对于网络关系的认知的变化。虽然知情人可能是外在的观察者，但是他们也是其所报告的网络的参与者。为了创建一个 CSS 数据库，研究者要求每一位参与者报告他/她对于每一个对偶网络关系的主观认识。例如，使用一个涵盖所有双方行为参与人的名单，去核对所指关系的感知性存在。对于某一个具有 N 个行为人的网络，一类关系的完整 CSS 数据库含有 N 个认知图，这些认知图标明了每一个参与者—知情者视角下关系连接点的显现与缺失。鉴于可能需要大量的时间和回忆的负担——每一个知情人需对具体的有向对偶关系进行 N^2-N 次判断，CSS 数据收集工作常常限于非常小规模的网络，或者一个有限的任务范围(例如，"谁是 X 的三个最好的朋友?""Y 最信任哪五个人?")。举例来讲，Krackhardt(1987)收集了 21 个管理雇员中的两种类型关系的 CSS 数据——建议寻求和友谊，得到了两套各 21 个认知图，每套展现了 $(21)^2-21=420$ 个有向联系。每个自我中心网络都是一个 CSS 特例，其中知情人只描述所观察到的他/她的直接联系人之间的关系，而不是与所有其他网络成员间的联系。

CSS 数据库的两个普遍运用是：(1) 判断参与者对网络结构认知的一致性程度；(2) 通过与某些准则进行比较，评估所观察网络结构的认知偏差。一致性

是指两个或更多的知情人对社会关系判断或评估的同意程度。一个重要的理论问题是，具有相似个人特性（比如种族、性别或年龄）的行为人之间的一致性程度是否更高？或者共享关系连接点（或位于相似网络位置）的行为人，是否更有可能在网络关系的认识上达成一致？偏差指知情人通过某一准则的比较，达到对网络认知的准确性程度（更多关于知情人偏差的问题，请参看后面章节）。其中一个准则是所有其他参与者所认知的网络的集合。另一个准则是从参与人行为的直接观察中所获得的关系数据。对自填式问卷测量与关系行为测量之间的较大差别，网络分析者已经观察很长一段时间了，并且已提出解决这种差别的办法。虽然大部分网络数据是通过面对面或者电话访谈获得，但是大部分研究人员很少注意受访人或知情人的认知社会结构，而这种认知社会结构会塑造知情人的回答。分析者一般将所有的反映当做明确客观的社会事实。然而，一些学者已经寻求理解和测算基于受访者认知获得的信息与行为观察之间的差异来源（比如 Batchelder，2002；Carley & Krackhardt，1996；Casciaro，1998；Freeman，1992；Johnson & Orbach，2002；Koehly & Pattison，2005；Krackhardt，1987）。

认知社会结构也帮助研究者确认受访者自我填答时出现的系统性认知误差。Kumbasar，Romney 和 Batchelder（1994）访谈了 25 个计算机硬件工程师，构建了一个认知社会结构。每一个知情人都被提供给一个随机抽选的工程师的名字，然后被要求从整个 25 人工程师名单中标注一个人，他们认为这个人是随机所选工程师的朋友。这一过程对所有的名字重复执行，包括知情人自己的名字。然后利用代数矩阵方法（参见第 4 章），获得的 25 个认知图能够整合到一个横跨所有工程师一知情人的全幅网络。Kumbasar 等人（1994）将自我认为的朋友网络与全幅网络相比较，得出结论认为，个人倾向于认为更亲近他们自己网络表征的中心，而不是全幅网络的中心。另外，个人倾向于构建一个具有认知连贯性的网络，也就是说，他们将自己看成是由朋友所包围的中心，而这些朋友也有大量的朋友圈子。

知情人倾向于构建一个具有认知连贯性的网络形象，这一发现并不新奇。关于认知心理学的大量文献认为，当人们遭遇关系不平衡时，他们会经历心理

或情绪上的压力(Heider, 1958; Newcomb, 1961)。随后,面临不平衡关系的个人开始启动"平衡计划":他们改变他们的认知,以求获得这一关系的平衡性认知图像。然而,这种认知的调节主要发生在朋友关系间,尤其是亲密朋友关系间,而在非朋友关系间人们似乎不受这种认知的不平衡所影响。实际上,当个人之间没有直接联系时,认知的不平衡很少导致心理上的不平衡。因此,知情人虽然尽力保持他们的心理平衡,但是他们还是更易于在朋友间提供扭曲的和有偏差的信息,相反他们对于疏远关系的报告更少受到偏差和扭曲方面的质疑。

与这种传统认识形成对比的是,Krackhardt 和 Kilduff(1999)从 Taylor 和 Fiske(1978)的认知吝啬者模型中得到启示,认为知情者一般都缺乏与他们并不直接联系的远距离关系人的了解。通过动员平衡计划来填补信息空缺,知情者在报告远距离关系和亲密关系时都会存在偏差。但是,两种偏差的动因是不同的:对于亲密关系来说,平衡计划是用于获取心理和情绪上的平衡,但是对于远距离关系来说,平衡计划是用于弥补有关联系信息的缺失。Krackhardt 和 Kilduff(1999)调查了 4 家公司的员工,请每一个知情人士报告其同事间的朋友关系。他们发现了网络结构与认知平衡间的曲线关系:员工认为他们的亲密朋友圈和最边缘关系圈要比那些中等关系的人更易于平衡。其中一个主要的启示便是,近距离关系和远距离关系都容易产生高度信息偏差,这是下一节的论题。

信息偏差

我们将信息偏差定义为自我报告的行为与真实行为之间的不一致。对于言词和行为之间的鸿沟,社会科学家早已开始进行观察。在 20 世纪 30 年代早期,Richard LaPierre 和一对年轻的中国夫妇开车横穿美国,途中他们在 184 家餐馆吃过饭,在 66 家旅馆住过宿,并且只被一家旅馆拒绝过。旅程结束后,LaPierre 给所有的餐馆和旅馆等发信,问他们是否"接待过中国人"。结果,92%的回复是"否"(LaPierre, 1934)。拉皮尔因此总结,偏见态度的行为表现

是需要支撑条件的。在另一项研究中,Kronenfeld(1972)询问刚离开餐厅的人,男侍者穿着什么样的服装?餐厅又播放什么音乐?许多受访者对服装和音乐提供了详尽的信息,但实际上这几家餐厅根本就没有侍者或没有播放过音乐。

信息偏差的来源多种多样。许多受访人对于准确报告他们自己行为的要求,并不具备处理这种大量信息的认知能力(Bernard, Killworth, Sailer & Kronenfeld, 1984)。另外还有两个方面的原因:知情人在认知上将某一种归类形式强加至某一不归类的附属模型上(Freeman, 1992);知情人矫正他们的认知以便维持亲密朋友间的平衡关系(Krackhardt & Kilduff, 1999)。一些研究人员认为,信息偏差并非随机的噪声,恰恰某些受访知情人在准确提供行为信息上的确要比其他一些人强。例如,一些研究表明,对受访主题的熟悉往往会有助于受访者对事件的回忆(Romney & Weller, 1984; Romney, Weller & Batchelder, 1986)。相反,另一项研究发现,虽然对论题知识渊博的受访者遗忘较少,但他们也容易同时创造出一些诸如并不存在的关系成员之类的错误(Freeman, Romney & Freeman, 1987)。还有分析发现,所有的受访知情人士都倾向于将自己视为更加靠近网络关系的中心,而并非如其他网络成员所报告的那样。因此,对某一网络组别层面的报告也许要比个人层面的报告来得更准确些(Kumbasar et al. , 1994)。最近几项关于自我中心网络的研究都对信息偏差问题进行了考量(例如:Brewer, 2000; Feld & Carter, 2002; Johnson & Orbach, 2002; White & Watkins, 2000)。

对于将自填式问卷得到的数据作为网络关系人的真实行为数据进行研究的社会网络研究者来说,也许最引人注目的消息就是 H. Russell Bernard 和他的同事们所进行的项目(Bernard & Killworth, 1977; Bernard, Killwoth & Sailer, 1981; Bernard et al. , 1984)。他们构建并比较了 7 套配对的交流网络数据库:由参与人自述建构的认知数据;由机器监控系统或"客观"观察者记录所建构的行为数据。这些数据库系统包括:两组华盛顿特区的聋人 TTY(文本电话系统,专为听力或语言不便者使用)使用者;弗吉尼亚、宾夕法尼亚和俄亥俄无线电协会的业余使用者;社会科学研究所的职员;西弗吉尼亚大学技术教

育的一个研究生项目；大学学生联谊会驻地；一组电子信息交换系统(EIES)使用者。例如，EIES使用者记录自述数据(认知网络)和通过 EIES 电子监控系统所得数据(行为数据)。所有 7 组双方交流网络数据包括："受访人与谁联系交流?""他们多长时间联系一次?""他们联系多长时间了?"通过对这些认知数据和行为数据进行比较，Bernard 等人(1981:17)发现大约一半的自我报告知情人在某些方面是存在问题的，并得出如下结论：

如果对其要求任何可接受的精确度，人们实际上并不能够说出他们在过去特定的一段时间内与谁交谈过……我们现在确信，关于交流的认知数据并不能够代表同等行为的数据。

其他一些学者认为，情况也并非像 Bernard 和他的小组所描述的那样毫无希望。Knoke 和 Kuklinski(1982:32)对"客观"观察者的准确性和中立性表示怀疑。他们提出如下一些问题，"在什么程度上这些中立的观察者打扰了正在进行的行为?""他能够多精确地记录双方的交流?""在非观察期他们之间的交流如何?"Bernard 等人(1984)后来调整了他们的例子，承认尽管在认知数据方面存在一些系统性扭曲，但通过一些改进因素可以减少信息偏差，比如让知情人更熟悉文化背景，或由访谈者给受访者提供相关回忆提示和帮助。他们建议"信息的不准确性不应该导致抱怨或绝望，而应该引至一个尚未开垦的研究新领域"(Bernard et al.，1984:513)。

信息偏差常常以其他系统形式发生。持偏差观点的受访者将自己视为更趋网络中心(Kumbasar et al.，1994)，而忘却边缘行为人，错误地回忆主要网络成员(Freeman et al.，1987)，在认知上矫正他们的认识以达成他们与或近或远的朋友间网络关系的平衡(Krackhardt & Kilduff，1999)。信息偏差也往往由不准确的回顾所致(Bernard et al.，1981；Bernard et al.，1984)。不能准确回忆的知情者容易犯两类错误：(1) 忘记他们与谁交往过；(2) 错误地回忆实际上从未发生的事(Freeman et al.，1987)。而且，遗忘和错误回忆并不是随机发生的。知情人倾向于遗忘并不频繁的参与者，错误回忆频繁的参与者。Freeman 等人(1987)记录了参加 9 个座谈会的大学教师和学生。最后两个会议结束后 5 天，研究者调查了 17 位参加过这两次会议的知情人和 16 位没有参加这两次会

议但是参加了至少两次之前会议的知情人。研究人员问他们参加和缺席会议的情况。这些知情人在 272 个选择机会(17 × 16 = 272)中犯了 141 个错误(115 个遗忘和 26 个错误回忆)。52%的错误率非常接近 Bernard 等人(1981)的估算。Bernard 等人估计过半的知情人自我报告是不准确的。座谈会知情人在对其他参与人的了解上,差异较大,而这就影响到他们的错误率(Freeman et al.，1987:312)。与知之甚少的知情人(他们只在座谈会中与其他人交流过)相比,知识丰富的知情人(他们对其他人在许多方面都有所了解)更少遗忘,但更多地犯回顾性错误。因此,与通常的认识相反,知识丰富的人实际上更容易错误回忆。然而,这些知识丰富容易错误回忆的人却是会议的经常参与者。由于社会研究的主要关注点是如何测量各种重复行为的模式,因此就收集这些数据而言,这些知识丰富的人倒是很好的知情人(Freeman et al.，1987)。

知情人之间的高度一致常常意味着低误差和高信息效度。因此,Romney 等人(1986)提出一个共识模型,将知情人的能力(指知情人了解研究客体的程度)和知情的统一性、有效性联系起来。一个核心观点是:任何两个知情人答案之间的一致性,是每一个答案与真相相关联程度的函数。也就是说,相对于无效的回应,有效的回应更容易经常得到同样的答案。研究者要求知情人回应一系列问题,而对于这些问题只有研究者本人才有标准答案。他们用所答正确数除以答案总数的比例来评价每一个知情人答案的“真实”有效性水平,用共识模型来评价其一致性有效水平。他们发现知情人的真实性准确度与他们的一致性准确度高度相关,因此他们也认为一致性的答案往往等同于有效的答案。

总之,与 Bernard 等人(1981)宣称的知情人自我报告总是错误地反映事实真相这一结论相反,对信息偏差的研究为如何更合理利用知情人提供了有益的建议。知识丰富的知情人能提供有关长期重复模式的无偏数据,同时他们也能够提供一致的答案,而这正意味着更高的有效性。个体认知相比于系统对所有网络行为人的综合认知而言,更偏向于将自己视为网络的中心。因此,系统认知要比个体认知更能准确把握一个人在网络中的定位。另外,自我往往会通过“平衡计划”对亲密朋友和较疏远的联系人产生认知偏差。那些利用自我中心

关系法来收集网络数据的研究者，在分析亲密关系和较疏远关系时，应当采取一些矫正措施。

信度

从一般意义上来讲，信度是指特定工具用于重复测量同一事物时，每次都得到同样结果的一种程度。其中，有好几种测量信度的方法，包括观察者间信度、再测法信度、内在一致性信度（包含折半信度和 Cronbach α 信度）（Neuman，2000：178—185）。社会网络研究者常常使用再测法来判断知情人的信度。例如，对知情人提出或剔除讨论重要问题的联系人名册，再测法会在以后的某一时间再次就同一问题调查该知情人。比较前后两次反应的一致性程度，即可得知每一个知情人的信度。当两次调查中知情人都提供了完全相同的核心讨论圈名单时，即这两次测量能够完美对应时，就表示完全可信。相反，在一个完全不对应的测量中，知情人选择了与上次完全不同的名单，那就意味着完全不可信。在操作再测法时，一个关键的问题是两次测量之间的间隔。两次测量之间的不同答案，也许反映的是个人网络关系的真实变化。实际上，人们不停地发展新的人际关系，他们之间可以进行表达性的或实质性的帮助，比如讨论重要问题、搬家、照看小孩或者借钱等。方法论者认为，这种颠覆性问题可以通过缩短两次测量之间的间隔来减弱其影响（Brewer，2000）。间隔时间越短，个人网络产生的实际变化可能就越小，因此这些变化对信度测量产生扰乱的可能性也就越小。

一个简单的但经常使用到的再测法信度测量是 Jaccard 系数，它指排除共同负面配对之后的同意部分。它是这样计算的：用知情人两次测量都提名的人数，除以两次测量中提名的总数①（Brewer，2000）。Jaccard 系数值从 0.00 至 1.00，表示零可信度和完全可信度。下面是一个含有 35 个姓名的虚拟名单，结果显示在交叉表中，单元频率按照 A—D 来表示。

---

① 两次都提名的人算一个人名。——译者注

第一轮访谈	第二轮访谈	
	提名人数	未提名人数
提名人数	A = 12	B = 5
未提名人数	C = 3	D = 15

在第一轮访谈中，知情人选择了17个提名；在第二轮中选择了15个提名；在这两轮中共有12个人被同时提名。因此，Jaccard系数 = A/(A + B + C) = 12/(12+5+3) = 0.60。一个小组中的每一个知情人可能具有不同的信度，因此一个小组级别的Jaccard系数便是所有个人系数的均值或中位数。

知情人信度常常与有效性相关联(Romney & Weller, 1984)。假如知情人的自我认知报告只有一部分反映了他们的真实行为，那么知情人之间的一致性就往往能够表明知情人回忆的有效性水平。例如，针对网球比赛的问题，网球运动员应当比非网球运动员表现出更多的一致性。知识丰富的运动员也许能够完全一致地回答这些问题，而没有相关知识的运动员则会给出五花八门的答案。从这个中心观点出发，Romney和Weller(1984)再次分析了Bernard等人的数据库，用于确认那些有效可信的"好的"知情人，和那些无效不可信的"差的"知情人。他们发现，个人的信度与他/她的效度是正相关的。个人的信度通过个人回顾与组别整体回顾之间的相互关系来测量，效度则根据个人回顾与组别整体数据的一致程度来计算。另外，高信度的人比低信度的人更倾向于与其自我报告有更高的相关性。因此，主要的发现是，可以通过信度预测效度，具有高效度的知情人之间应该要比低信度的知情人之间，具有更多的相似反应。而且，知情人应当由他们的信度来加权：那些具有更高信度知情人的答案应该比那些低信度知情人的答案赋予更高的权重。仅就高信度人群这点来说，Bernard等人的数据中的知情人提供了有效的自我报告，这从他们的回忆数据和综合观察互动数据的高度一致性能看出来(Romney & Weller, 1984:75)。

在一定程度上，知情者中忘却某些联系是不可避免的。因此，相对于比较自由回忆法，可以采取某些帮助回忆措施——比如提供线索或提示的名单——它们的可靠性问题更少。比较自由回忆法是指知情人在没有任何帮助的情景

下进行提名。例如，Jones 和 Fischer(1978)对某种具体网络关系，比如“借钱”、“讨论工作有关的事务”，访谈了 86 个人。第一轮访谈中，受访者自由地回忆那些关系中的个人。第二轮，研究人员给每位受访者一个名单，并且重复了同样的问题。结果第二轮比第一轮多得到 34%的人员名单，而且第一轮提名的 27%在第二轮访谈中并没有被提及。

在另一项网络信度研究中，Brewer 和 Webster(1999)调查了 217 位住在一个大学宿舍楼中的人员。第一轮调查要求他们写下居住在这栋大楼的所有朋友的姓名。第二轮则为每个受访人提供了住在本楼的人员花名册，然后请受访人再次确认他们认识的朋友。利用花名册的提示，受访者平均多提出了 20%的朋友名单。而先前忘记的名字，多是关系比较疏远的人，这也表明如果没有提示和线索，受访者一般会遗忘与他们关系相对疏远的对象。其中一个推论是，当一个研究人员将在两次独立的访谈中获得的自由回忆数据和基于帮助而得到的数据进行比较时，会出现结果非常不可信的情形。在这种情况下，低信度只意味着更高的效度：知情人在一个访谈中比在另一个访谈中得到更完整的提名。当然，信度并不保证效度。一个知情人在两次访谈中也许提名完全相同，因此获得完美的信度，然而每轮访谈都可能因同时忽略了同样的名字而产生了同样的偏差。不过，高效度的网络反应一般来说具有高信度。也就是说，假如一个受访者能够总是根据某一提名法则列举同样的朋友名单，那么这个受访人也很可能在不同的访谈时间里选出一致的朋友圈子名单。

由于社会网络数据的特有属性，尤其是在自我中心数据分析中，受访者的信度和效度测量与传统的其他种类数据的测量差异非常大(Marsden，1993)。从较小且封闭的社会中抽取的自我中心样本中，知情主体常常提出许多与某一具体关系有关的相同的关系客体。知情人之间在选择相似关系客体上所体现的高度相关性，意味着知情人的高度可信。同样地，知情人对他/她的关系客体特征(如年龄、性别、教育背景、社会地位等)的描述，与关系客体对这些特征的自我报告二者之间的高度相关，也意味着知情人的高度有效性(这个解释假定所有关系客体的特征都可从其自我报告中获知，从而用于作为比较知情人自我报告的事实数据)。

研究者对肯尼亚乡村妇女间的避孕话题网络进行过一个专门调查，揭示了这种自我中心网络方法的信度和效度问题（White & Watkins, 2000）。研究者在1994年12月访问了925位妇女，一年后再次访问了其中的743位。在两次访问中，这些妇女都被问及同样的问题："你与谁商讨有关节育问题？"受访者随后报告其中4个随机抽选出来的关系客体的年龄、教育背景、收入状况以及节育措施的使用情况。研究人员也询问每一位关系客体自己的年龄、教育背景、收入状况和节育措施情况。White 和 Watkins 发现，在受访者自我报告和关系客体自我报告之间，节育措施情况存在最明显的差异。相反，在一些可直接观察到的事项上则存在较高的一致性，比如"朋友家的房子是否有一个金属屋顶"或"她家里是否有套沙发"。而且，由受访者（提出了同样的关系客体）所提的关于关系客体的报告的相关度只有 $r = 0.31$，只比受访者自我报告与关系客体自我报告的平均相关度(0.25)高一点点。这些发现表明，受访者对于关系客体节育措施的了解是既不可信也非有效的。第一轮由受访者提出的关系客体只有18%在第二轮中提到，这也表明了或者是低信度，或者是在受观察年份期间中关系网络的不稳定。

数据缺失

社会网络研究对数据缺失尤其敏感。在自我中心网络数据收集过程中，一个拥有 N 个关系客体的关系主体被要求报告 C_N^2 个非定向的对偶关系。具体来说，$C_N^2 = \frac{N!}{2! \times (N-2)!}$，其中 $N!$（读做"N 的阶乘"）表示从1至 N 的所有正整数的乘积。例如，假如关系主体有5个关系客体，则 $C_N^2 = \frac{5!}{2! \times (5-2)!} = \frac{(1 \times 2 \times 3 \times 4 \times 5)}{(1 \times 2) \times (1 \times 2 \times 3)} = \frac{120}{12} = 10$；因此，关系主体必须对这10种非定向关系给出有关信息。对于定向关系，关系客体间的连接点数量就是上面的两倍：$2C_N^2$；因此，在上述例子中，关系主体应当报告20种定向关系点。自我中心网络关系的关系应答率 R，等于所报告关系点的数量除以关系客体间所有可能

的对偶关系数量。比如，关系主体报告了10种非定向关系中的8个，那么$R=0.80$，或者80%；假如关系主体漏报了20种定向关系中的6个，那么$R=0.70$，或者70%。

要计算一个完全社会网络的应答率是比较复杂的一件事。一个完全的社会网络由网络中N个行为人配对所组成的所有双边关系所构成。对于非定向关系来说，R相对就简单些，因为当测量信度较高时，由双边关系中的一方来报告便足够了。例如，为了测算行为人A和B之间的友谊，由任何一方知情者所提供的信息即可判断这种朋友关系的存在与否。也就是说，除非A和B的报告都缺失，否则我们就可以凭借一方的单独报告来判定友谊的存续。一般来说，对于一个含有N个行为人的非定向关系网络，当其中有M个行为人没有关系客体报告时，某一具体关系的应答率的计算如下所示：

$$R\begin{cases}=100\% & \text{当 } M=0 \text{ 或 } M=1\\ =\left(1-\dfrac{C_M^2}{C_N^2}\right)\times 100\% & \text{当 } 1<M<N\\ =0 & \text{当 } M=N\end{cases}$$

例如，在一个含有5个行为人的网络中，节点应答率和关系应答率依关系节点缺失数量的不同而有所变化，具体见下表：

缺失节点数	节点应答率	关系应答率
0	100%	100%
1	80%	100%
2	60%	90%
3	40%	70%
4	20%	40%
5	0	0

为了方便阐述，现假设网络行为人分别为A、B、C、D、E。在这5个行为人中的10种非定向对偶关系是AB、AC、AD、AE、BC、BD、BE、CD、CE、DE。假如行为人A没有报告他/她的关系，关于A的这些对偶节点依然能够从其他4个人的报告中获得。因此，尽管缺失了1个节点的报告，关系应答率仍然

是100%。当节点缺失数量为2至4（$1<M<N$）时，关系应答率是$\left(1-\frac{C_M^2}{C_N^2}\right)\times 100\%$。比如，假如有3个节点(A、B、C)未报告他们与其他人的关系时，关系应答率为70%$\left[\left(1-\frac{C_3^2}{C_5^2}\right)\times 100=70\%\right]$。在这种情况下，3对非定向性关系缺失了(AB、AC、BC)，但是其他7对关系仍然可以由其他剩下的人报告。假如没有任何行为人提供信息，则节点应答率和关系应答率都降为0。因为节点应答率的计算公式为$\left(1-\frac{M}{N}\right)\times 100$，而且$\frac{M}{N}$总是要大于$\frac{C_M^2}{C_N^2}$（有关数学证明可以向本书作者索要），因此在节点缺失报告的每一个层级上，非定向关系网络的关系应答率总是会高于节点应答率。

相反，缺失的节点信息对于定向网络的关系应答率具有重大影响。不对称性存在于多种网络关系中，例如提供建议、信任、喜爱。行为人A和B可以构成两种定向性关系：AB，表示A所报告的其与B的关系；BA，表示B所报告的其与A的关系。因此，每一个节点的缺失，都会导致该节点所指向的所有关系信息的缺失。假定一个含有N个行为人的网络缺失了M个节点，那么该定向网络的关系应答率为：

$$R\begin{cases}=100\% & \text{当} M=0\\ =\left(1-\dfrac{M\times(N-1)}{2\times C_N^2}\right)\times 100\% & \text{当} 0<M<N\\ =0 & \text{当} M=N\end{cases}$$

说明如下：现假定有一个5人的关系网络，它含有$(5^2-5)=20$种定向对偶关系：AB、BA、AC、CA、AD、DA、AE、EA、BC、CB、BD、DB、BE、EB、CD、DC、CE、EC、DE和ED。如果每个行为人都报告其与他人的关系，那我们就得到一个100%关系应答率。假如有一个点(比如A)缺失，那么所有A的报告也会缺失，导致这20对关系有4对(AB、AC、AD、AE)未报告，因此关系应答率就只有80%。在定向网络中的每一个点都必须评价与其他点的关系，因此M个缺失点(当$0<M<N$)将导致$M\times(N-1)$个关系的缺失。当然，当没

有一个点报告关系时($M=N$)，节点应答率和关系应答率都等于0。

利用一点代数推理，我们现在证明在定向网络中关系应答率总是等于节点应答率。假设我们有一个含 N 个节点、M 个缺失点的定向网络，那么节点应答率是 $\left(1-\frac{M}{N}\right)$，而关系应答率为 $\left(1-\frac{M\times(N-1)}{2\times C_N^2}\right)$。具体来说，$\frac{M\times(N-1)}{2\times C_N^2}=M\times(N-1)\div\left(2\times\frac{N!}{2!\times(N-2)!}\right)=\frac{M\times(N-1)}{N\times(N-1)}=\frac{M}{N}$。因此，在有5个节点的定向网络中，当没有节点缺失时，节点应答率和关系应答率都是100%；当1个节点缺失时，这两种应答率都为80%；当2个节点缺失时，为60%；当3个节点缺失时，为40%；当4个节点缺失时，为20%；当所有节点都缺失时，为0。

在前面的论述中，我们认为，在非定向网络中，对偶关系中的某个成员所进行的报告可以视为是对该对偶关系的可信测量。但学者一般都认为，这一观点必须谨慎采用，因为经常发生单向关系(Stork & Richards, 1992)。例如，在一个交流网络中，A报告说与B进行交谈，但B报告说根本未与A交流。这种矛盾对下列做法提出了质疑，即仅从未缺席方提供的信息，来构建行为应答方与行为缺席方的关系。Stork和Richards(1992)建议，关系双方特征的相似性，比如他们的年龄、性别和教育背景等，可作为认定他们对偶关系的一个很好的指标。另外，网络信度——由双方关系成员共同描述的所有对偶关系的比例——表明了一个单独的报告是在何种程度上揭示调查应答者与未应答者之间的联系。当网络信度高，并且关系双方具有相同的或相似的个人特征时，一方的报告可以作为他们关系的一个可信代表。Kossinets(2003)赞同Stork和Richards的方法，即当调查未应答者数量很少时，可以根据应答者的报告来构建对偶关系网络。

自我中心网络研究对数据缺失比较敏感，因为人们常常疏于描述关系客体间的联系。在研究1985年GSS讨论网络时，Burt(1987)的报告指出，GSS自我中心网络研究中的数据缺失并不严重：共有1534个调查回应者，他们列举了4483个讨论对象，其中只有66个应答者提供的网络数据不完整，涉及195个讨

论对象。Burt 比较了两种讨论对象——一种具有全部网络关系数据，另一种则只有部分网络关系数据。他发现前者关系客体的35%都与其他关系客体关系不紧密，这正如调查应答者所报告的那样，而后者的这个值则升至59%。这其中主要的含义是，在一个自我中心关系网络中，关系客体间所缺失的关系一般都是弱关系。关系主体似乎更倾向于报告那些与其他关系客体存在联系的客体关系，而不是那些看起来比较孤立的客体关系。这一发现也解释了与其他研究方法结论相反的一个现象：受过大学教育的 GSS 调查回应者，相比教育程度较低的回应者，往往会提供更加不全面的自我中心网络关系数据。大学毕业生倾向于列举那些由许多弱联系的关系客体所构成的自我中心网络，反之也将不会更多地报告有关客体关系的数据缺失。

由于没有万无一失的事后补救办法来解决数据缺失问题，因此这个问题的解决之道只能是让更多的调查对象了解参与的重要性(Knoke & Kuklinski，1982：35)。为了获得较高的参与程度，比如有些网络研究达到了90%的应答率，研究人员常常要付出更多的辛苦，采取不同的说服技巧，比如个人书信、电话联络以及钱财激励等。数据缺失问题这一难题并非仅仅存在于使用调查方法进行的社会网络研究中，档案研究也容易受数据缺失的影响。一般人会想当然地认为，档案数据的完整性是受数据存储管理操作的影响的。然而，在某类数据的可得性不变的情况下，编码人员的能力与数据搜索软件的效率也同样能对获取信息的数量和质量产生重大影响。训练较差的编码员、设计糟糕的软件，都绝对会导致规模较大的数据缺失。

第 4 章 | 网络分析基本方法

这一章我们将讨论社会网络分析的基本方法，对重要的传统论题和新近的发展也会给予同等的关注。其中，我们尤其着重论述密度、中心度、内聚、结构等价、聚类、多维度标度、块模型。由于网络数据描述总是在社会网络分析之前，因此这部分自然也就从描述两种独立但联系紧密的展示和分析社会网络数据的方法开始：图表和矩阵。图表展示的是社会网络的可视性，而矩阵则利用数学上的代数来描述网络关系。虽然社会网络学者可以自由地选择运用图表或矩阵模式来描述他们获得的数据，但是这两种方法各有优劣。图表可以为读者提供有关网络结构的更加直观的表述，但缺乏数学上的可操作性。与之对应的是，虽然矩阵看上去界面不够亲近读者，但他们非常便于对社会网络数据进行复杂的数学和计算机分析。网络分析者必须熟练运用这两种日常分析工具，以便对关系数据的重要特征进行概念化和测量。

图表

Jacob Moreno(1934)是社会计量方法的先驱，他重视建立社会关系图，即二维图，用于展现某一有界社会系统内行为人之间的关系，比如一个小学班级。点和线——社会关系图中的两个基本元素——也确定了图的基本构成，从而也从图论中引入了相关概念和定理(Harary, 1969)。在社会关系图和图表中，行为人是由 N 个点(也称为节点或顶点)来代表的，常常用标识名、字母或数字来表示。两点之间划出的线(也称为弧或边)，表示一种关系或连接；线的缺失，意

味着两个行为人之间不存在直接关系。如果两点之间有一条线，那么这两点就是邻接的。如果某一个点是定义某条线的两点之中的一点，那么这个点就附属于这条线；反之亦然。在网络关系量表中，点的精确位置和线的长度是可以随意的，虽然某些版本要比另外的版本更清晰。建构一个富有洞察力的社会关系图，和进行一项科学活动一样，是一种艺术行为。

假如一条线没有箭头，那么就表示这种关系是非定向性的或双向的（比如同事关系）。假如关系是从一个行为人指向另一行为人，那结果就是一个定向图表或有向图。形式上，一个有向图是点的有限集合，也是有序配对(a, b)的集合，其中点a是线的起点，点b是线的终点。单向箭头表示一种定向关系，方向由箭尾的行为人指向箭头的行为人（例如，提供建议）。双向箭头线条表示两个定向性联系，从每一个点到另一个点，表达的是一种互惠关系或相互关系（比如，每个行为人都选择对方作为自己的"亲密朋友"）。一个替代模式是使用两个单向箭头，每个箭头都指向对方。一些社会关系图也通过线条粗细的变化，或者不同种类破折号的使用，以便更形象地表达双方关系的力度、密度或者频率。

为了阐述一个社会网络结构的图表，我们利用了 Feldman-Savelsberg 等人(2005)对喀麦隆妇女同乡会组织的网络研究。为了便于分析集体记忆是如何影响妇女们对生育的讨论，Feldman-Savelsberg 等人(2005)访谈了喀麦隆首都雅温德的 6 个妇女协会中的 156 个成员。他们的深度访谈涉及妇女社会网络的一些问题，比如"请根据下列标准对你和你所在协会的其他妇女的联系力度进行排序：(1)知己；(2)朋友；(3)认识；(4)完全陌生"。由于每一个妇女仅被问及她与同一协会其他妇女的关系排序，因此对这 6 个协会都分别构建了独立的网络结构。为了简单阐述，我们选取了第六协会的网络结构，它有 6 个协会成员。

图 4.1 表示第六协会成员间的社会关系图。线条的粗细表示 3 种层级的选择，没有连线表示两人之间完全陌生。在这 6 个成员间的 23 种定向关系中，只有 15 对关系中的 4 对是互相同意的（即两个妇女都选择了同样的层级）：妇女 1 和 2 互相视为知己，妇女 3 和 5 是朋友，妇女 2 和 5 仅仅认识，而妇女 5 和 6 则

完全陌生。其他 11 对则对她们的关系有不同的分级。例如,妇女 6 将妇女 2 视为知己,但妇女 2 认为仅认识妇女 6 而已。虽然社会关系图可以提供一个直观的网络结构图,但是在展示由几十个甚至上百个行为人组成的大型网络方面,它们在视觉效果上就常显得很难看了。我们可以看出,仅仅 6 个行为人和 23 种定向关系,图表就显得非常复杂了。相比较而言,一个矩阵就可以轻而易举地表示几十个行为人的网络关系。

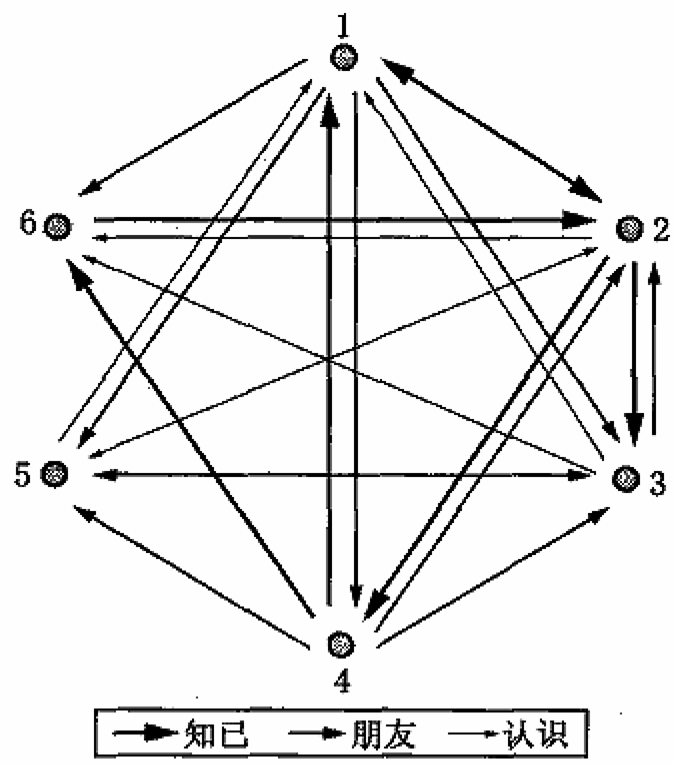


图 4.1 喀麦隆首都雅温德第六妇女同乡会的社会网络

图 4.2 是一个包含 7 个虚拟行为人的有向图,表示的是提供建议的社会关系。通过这张图可以阐述几个基本的概念。子图,是指在一张大图中点和线的子集。Tom, Betty, Jane 3 人组成的三角关系便是一个子图。行程是附属的点和线交替的过程,在这个过程中,每个节点附属于它前后的线。行程的长度是它所包含的线的数量。行程 Jane—Tom—Betty—Harry—Dick—Sally 的长度是 5。轨迹是由不同的线构成的行程,即使其中一些点也许被访问过不止一次,例如 Dick—Sally—Harry—Tom—Betty—Harry—Dick 便是一个长度为 6 的轨迹。

由于这个轨迹的起始和终止都在同一个点,因此它也被称为回路。Wasserman 和 Faust(1994:107—108)讨论了封闭行程、旅程和回路等相关概念。

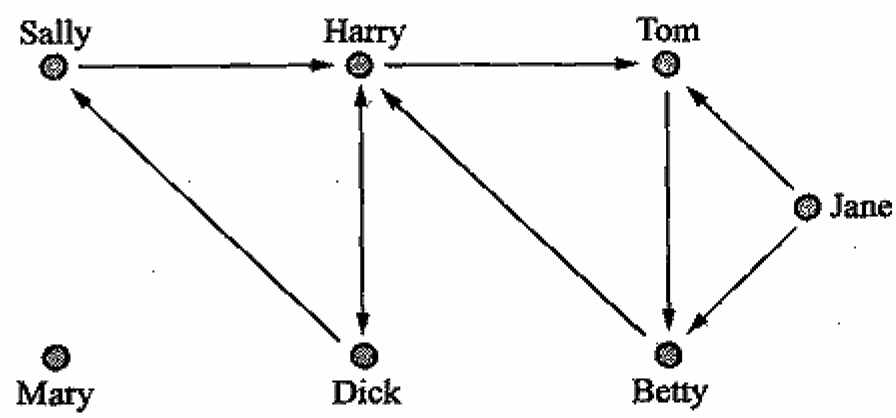


图 4.2　由 7 个行为人组成的有向社会网络

在网络分析中，一个重要概念是路径。它被定义为一个经由完全不同的节点和线的行程，其中点和线只能经过一次。在有向图中，一条路径必须经由不同的节点，而且其中的每条线都是由第一个节点到第二个节点。换句话说，所有的箭头都是朝着一个方向的。因此，Betty—Harry—Sally 不是一条路径，但Betty—Harry—Dick—Sally 是一条路径(见图 4.2)。路径的长度(即其行程中所经线的条数)被称为径长。在两个点之间也许有好几条路径，但是最短的路径被称为测地线。从 Jane 到 Sally 的两条路径中，测地线 Jane—Betty—Harry—Dick—Sally 的径长为 4，它短于另一条长度为 5 的路径(Jane—Tom—Betty—Harry—Dick—Sally)。那从 Dick 到 Tom 的测地线是多少呢？

如果两个节点之间存在至少一条任何长度的路径，那么这两点就是可及的。Tom 对 Sally 来说是可及的；虽然对 Sally 来说，Jane 是不可及的，但是对 Jane 来说 Sally 是可及的。

如果每对节点之间存在路径，那么这个图表就是连接状态的；但如果至少有一对节点之间没有路径，那这个图标就是非连接状态的。如果一个点没有与任何其他点有线连接，那这个点就是孤立的。Mary 明显是孤立于其他 6 个人的，因此这个图也是处于非连接状态的。假如我们忽略 Mary，那剩余的 6 人关系图就展示了三个层次的连接：

强连接：每两个点都由直接的路径双向连接(例如，从 a 到 b 并且还从 b 到 a)；

单向连接：每两点之间都从一个方向（而非另外方向）由一条路径相连接；

弱连接：所有两点之间有连线，但不考虑其连接方向。

很明显，这个6人的有向图只是个弱连接图。然而，在大网络之中的几个小网络子集的连接强度要比整个网络图强。图表组成部分是构成一个连接状态图表的极大化子图，而一个强连接图组成部分就是构成强连接图表的极大化子图。子图 Sally—Harry—Tom—Betty 是一个单项连接部分，因为不同长度的定向性路径单方向地连接了每一对点。这个组成部分是极大化的，因为一旦加入 Jane，就将破坏这种单向联系。那为什么 Jane—Tom—Betty 这一更小的子图不属于强连接部分或单向连接部分呢？

在一个连接图中，如果去除一个点，将导致该图不能连接（即会创建两个或多个组成部分），那这个点就是一个分界点。假如将 Harry 从图 4.2 中移除，那这个六点图就会分成两个小的组成部分（Sally—Dick 和 Tom—Betty—Jane）。同样的，如果去除一条线，会将一个图表分成两个或多个组成部分，那这条线就称为桥。请问这张图中包含桥么？为什么？

相对于存在很多维持信息和资源流的多余路径的网络，有分界点和桥的网络更容易受到破坏。例如，恐怖基地的秘密网络有可能因为一个关键人物被抓或被杀而陷入瘫痪状态（Borgatti，2003；Tsvetovat & Carley，2005）。

矩阵

网络关系的代数形式可以表达社会关系图中所有的量化信息，和直观的图表形式相比，代数形式能够做更多的分析。可供数学分析的社会网络数据的最基本形式是列表，称为社会矩阵，也称为邻接矩阵，是一个由行和列所构成的数值方列阵。例如，大写黑体的表达式 $\mathbf{X}(N, N)$ 表示社会矩阵 **X**，它有 N 个行和列，而且其 N^2 个单元表示 N 个社会行为人之间的各种关系。行为人的顺序或定位可以通过矩阵的行和列来一一对应，利用下标字母 i 和 j（其

值从 1 至 N)，相应表示特定的行列位置。大体上，一个单位格的值表示为 x_{ij}；例如，x_{35}表示一个社会矩阵中第 3 行第 5 列的值。在绝大多数社会网络的实际运用中，矩阵主对角线(即第 i 行第 i 列所组成的单位格)是没有意义的，例如，某个人不可能是他/她自己的好朋友，因此在数据分析中是可以被忽略的。因此，在一个含有 N 个行为人的社会矩阵里，拥有唯一值的单元格的最大值是 $N^2 - N$。

在一个社会矩阵单元格的数值测量矩阵中，特定的行和列对应一对行为人之间的特定关系。按照惯例，在定向关系中，行中的行为人属于起点或发送者，而列中的行为人属于终点或关系的接受者。最基本的测量就是关系的存在或缺失，分别由 1 和 0 这两个二进制值表示。因此，二进制矩阵数 $x_{ij} = 1$ 表示行为人 i 发送了一个关系给行为人 j；如果 $x_{ij} = 0$，则表示由 i 至 j 没有关系。由于在有向图中，关系并不局限于双向的或者互惠的性质，因此它们的矩阵图是不对称的。在非定向网络中，$x_{ij} = x_{ji}$，也就是说，起点 i 和终点 j 之间的关系值总是等同于起点 j 和终点 i 之间的关系值。这种矩阵是对称的。

社会矩阵也包含非二进制值，反映关系强度，比如联系的频率、关系的力度或者组织的大小。在这类赋值的图表中，单元格可以从 0 至对偶关系水平的极大值，或者包括小数甚至负数以表示反向关系，比如“不喜欢”或者“敌对”。

表 4.1 给出了喀麦隆雅温德地区第六妇女同乡会的矩阵图。这个矩阵是图 4.1 的代数形式。由于每一位妇女都对与其他 5 位妇女的关系力度进行打分，并且在双方评价时会有差异，因此这个矩阵是赋值和不对称的。行中的妇女对与其他妇女的关系进行评价，而列中的妇女是被打分的对象(行和列中的标注并不是矩阵的一部分，但是可以作为参考)。例如，妇女 1 对与妇女 3 的关系力度评分为 2(表示朋友关系)，但是妇女 3 对她与妇女 1 的关系的评分只有 1(表示认识)。图中的破折号表示主对角线上的值是没有意义的。表 4.2 是图 4.2 的一个二进制赋值矩阵。

表 4.1　喀麦隆雅温德地区第六妇女同乡会的社会网络矩阵

	W1	W2	W3	W4	W5	W6
W1	—	3	2	2	2	2
W2	3	—	3	3	1	1
W3	1	2	—	0	2	1
W4	3	2	2	—	2	3
W5	1	1	2	0	—	0
W6	0	3	0	0	0	—

注：0 = 完全陌生；1 = 认识；2 = 朋友；3 = 知己。

表 4.2　7 人有向社会网络(图 4.2)的矩阵图

	Sally	Harry	Dick	Tom	Betty	Jane	Mary
Sally	—	1	0	0	0	0	0
Harry	0	—	1	1	0	0	0
Dick	1	1	—	0	0	0	0
Tom	0	0	0	—	1	0	0
Betty	0	1	0	0	—	0	0
Jane	0	0	0	1	1	—	0
Mary	0	0	0	0	0	0	—

有时，网络研究人员使用非正方形矩阵表示行为人特征或者他们对某一事件的参与。数学符号 $Z(N, M)$ 表示一个长方形矩阵 Z，其中 N 是行为人的数量，M 是特征、事件或地址的数量。例如，Freeman 和 Webster(1994)观察了 31 天内 43 个常规沙滩客，记录了与这些人有关的 353 个事件(比如野餐、游戏等)。他们建立了一个 43×353 的矩阵，其中第 i 行第 j 列的值为 1，表示第 i 个人参与了事件 j。如果研究人员收集了不止一种关系的网络数据，那么可以分开或综合构建并分析多个社会矩阵。表示 R 多元网络间具体关系的矩阵形式，只需要简单地多加个脚注，也就是 x_{ijk}，表示在第 k 种关系中第 i 行和第 j 列的单元值。

关系的测量方法

社会网络数据中的基本关系测量方法包括密度和中心度。在讨论它们之

前，我们回顾几个网络数据的基本原则，它们在向度和赋值这两个维度上有所区别。向度将社会网络数据分为无向关系和有向关系两种。在无向关系数据中，一对关系的起点和终点(例如，发送者和接受者，选择者和被选者)是没有区别的：如果 Dick 娶了 Jane，那 Jane 也同样嫁给了 Dick。其他种类的无向连接包括某一组织内成员关系，某一社会事件的共同参与，以及互相交流知识等。为了获取无向关系资料，研究者常常只需要问一方知情者，然后利用这个应答即可了解这个对偶连接。相反，当对偶关系的每一方都向对方启动一种关系时，有向关系发生了，结果导致 4 种可能的二进制组合：(1)A 给 B 一些东西，比如信息、建议和钱等，但 B 不给 A；(2)B 给 A，但 A 不给 B；(3)A 和 B 之间互相交换；(4)A 和 B 之间没有交易。有向数据经常收集到这样的情形：关系主体的自我报告认为存在某种网络关系，但这种关系与关系客体对其看法不一致。许多也许在理论上可以被归为无向关系的网络数据，在实际中应当视为是有向连接数据来测量。例如，为评估一项信任关系的存在，研究人员应当问关系双方是否信任对方。研究人员不应首先假定朋友关系是否总是互惠的，而应首先将他们视为是内在的有向关系：Ted 也许视 Sally 为朋友，但 Sally 可不认为 Ted 是她的朋友。正如第 3 章所讲到的，许多的认知社会结构网络关系具有固有的有向性，因为它们是来自于行为人对主观信仰、态度或者关系认识的报告，而这些很可能是不同于关系客体的主观看法的。即使一个关系在理论上可以归为无向类，但研究者通常还是得选择谨慎一点的做法，即请关系双方都提出各自的报告。在随后的分析中，研究者可以利用弱标准(比如只要一方报告认同，即认为关系存在)，或强标准(要求双方都必须确认该关系)来整合这种不同。一旦采用这一观点，那么在数据收集上就能使得分析者在供替代的标准假设之下验证他们结果的稳健性。

根据关系信息的可获得性，社会网络联系可以分为二进制的或赋值的量表。在二进制数据中，被试者只需要回应关系的存在(编码为 1)或缺失(编码为 0)。在赋值数据中，他们不但要表示关系的存在与否，还要对二者交往的强度或频率进行排序或量化。例如，青少年会被问及在去年与伙伴发生性关系的次数。因此，二进制网络数据只有两个值，0 或 1。但赋值关系数据则展示了一个

有序的或连续的梯度，数值可以从 0 至极大的可能值，甚至可以涵盖小数或分数。

将向度和值进行交叉列表，可以得到四种基本的社会网络数据类型：二进制无向、二进制有向、赋值无向、赋值有向。二进制无向数据需要从受访者或知情人那里获得的信息最少。访谈者只需要询问一方即可确认关系的存在与否。赋值无向关系数据需要一方来量化双方的关系。二进制有向测量只需要双方提及关系的存在与否。但是赋值有向数据需要最详尽的信息，即关于对偶关系重要性的两个报告。由于自我中心网络研究和全幅网络研究在它们的研究设计、抽样、数据分析及其解释等方面存在极大的差异，因此下面两个小节讨论适用于这 4 类网络数据的测量。

自我中心网络的测量 在自我中心网络研究中，知情人被要求报告与其有特别联系的关系客体。这种关系被明确表述在提供给被调查者的提名法则中。一个著名的例子就是美国综合社会调查，“讨论重要问题”的提名法，在第三部分有论述。没有独立的信息是从由 1985 年 GSS 被调查者所提名的任何关系客体那里收集的。这一小节主要集中介绍自我中心网络的密度和网络范围测量。

密度是指测量自我中心网络中 N 个客体之间互相联络的程度（忽略了关系主体 ego，因为本来所有的客体都与之直接联系）。现假定这种关系是无向性的二进制（有/无）关系，那关系密度（D）就等于所报告的客体间对偶联系（用 L 表示）数除以此种关系的极大可能数：

$$D=\frac{L}{C_N^2}$$

其中 $C_N^2=\frac{N!}{2!\times(N-2)!}$。例如，假设主体有 5 个关系客体，并且客体间只报告了 2 种关系，由于 $C_5^2=\frac{5!}{2!\times(5-2)!}=\frac{(1\times2\times3\times4\times5)}{(1\times2)\times(1\times2\times3)}=\frac{120}{12}=10$，那么这个社会网络的关系密度就是 $D=\frac{L}{C_N^2}=\frac{2}{10}=0.20$ 或 20%。

如果一个自我中心网络的联系是根据有向二进制来测量的，那么密度计算

公式就为：

$$D=\frac{L}{2\times C_N^2}$$

上述两种密度计算公式的第一个区别是分母不同。在同一 N 值的社会网络中，有向连接数是无向连接数的两倍。第二个区别是分子不同。虽然两个公式的分子都是反映对偶联系的真实呈现数量，但后者必须检查两个关系方向。在二进制无向网络数据中，每对连接方只计算一次，但在二进制有向网络数据中，每对连接方被计算两次，因为关系主体要报告每个客体是否首先与所有其他客体方进行联系。

接下来，假定网络数据是赋值的无向性关系，那么自我中心网络密度计算如下：

$$D=\frac{\sum L_w}{C_N^2}$$

在这个公式中，分子表示各种关系的所有值的和。此分子除以 N 个行为人间可能发生的所有关系数，便得到关系密度值，反映了主体之客体间关系的平均力度。假设某一研究项目要求主体对其 5 个关系客体（A、B、C、D 和 E）进行排序，标准如下：无关系（0）、认识（1）、朋友（2）、密友（3）。假如主体将 A 视为 B 的密友，将 D 视为 E 的朋友，而其余人之间没有关系，那么密度 $=(2+3)/10=0.5$，结果表明这个自我中心网络的 10 个关系呈现的平均力度或权重为 0.5。

最后一种密度测量是关于赋值有向关系数据的，公式如下：

$$D=\frac{\sum L_w}{2\times C_N^2}$$

它的分子包括所呈现关系的所有赋值的总和，而分母则包括 N 个行为人所在网络的所有可能有向关系数量的最大值，这是同样规模网络无向关系数的两倍。

二进制网络数据的密度测量值被限制在 0 和 1 之间，其中 0 表示关系客体

间没有任何联系，1 表示所有的客体间都存在直接互相联系。这种限制的存在是因为二进制网络密度的分子不可能大于分母。对于无向和有向二进制数据二者来说，双方实际关系的数量必然要小于或等于所有可能的对偶关系的最大数（当 L 仅表示关系的存在数量时，$L \leqslant C_N^2$；当 L 不仅考虑关系的存在，还考虑关系的方向时，$L \leqslant 2 \times C_N^2$）。然而，对赋值的自我中心网络数据来说，密度能够超过 1.00，因为对关系特征的赋值也许远远大于二进制数据的最大关系值 1。

网络范围在广义上被定义为一个行为人的网络与其他不同行为人的联系的程度。在实践中，研究者利用多种经验标准来测量网络范围，例如网络规模、密度和多样性（Campbell, Marsden & Hurlbert, 1986）。而且，自我中心网络多样性测量的是一个个体网络中客体社会特征异质的程度。依据客体特征的测量水平，自我中心网络密度可以通过连续变量的标准差来进行测量，或者通过分类变量（包括离散的或有序的）的定性变化指数来测量。

当一个客体特征按照连续变量来测量时，标准差反映的就是一个自我中心网络内特征的差异性。一般来说，对含有 N 个客体的第 i 个主体，其中第 j 个客体的特征 X_{ij} 是一个连续变量，那么标准差就是：

$$s_{X_i} = \sqrt{\frac{\sum_{j=1}^{N} (X_{ij} - \bar{X})^2}{N-1}}$$

所有主体标准差的平均值可以测量一个自我中心网络样本中特征的多样性。Marsden(1987)发现，1985 年 GSS 调查所得客体教育年限的平均标准差只有 1.37 年，这意味着在美国人的核心讨论圈中讨论对象都倾向于拥有相似的教育水平。

当客体特征是按照分类变量来测量时，那 IQV 就是用来测量自我中心网络的多样性。通常，对含有 N 个客体的第 i 个主体而言，其中所有客体都被分成 K 类离散的或有序的类别，那 IQV 等于

$$IQV_i = \frac{1 - \sum_{j=1}^{K} p_j^2}{(K-1)/K}$$

其中 p_j^2 是第 j 类中客体的比例。IQV 的标准值介于 0.00 和 1.00 之间，其中 0.00 表示所有的 N 种情形都包含在一个类别之中，而 1.00 表示所有的客体都均匀分布在 K 个类别内。例如，假如某主体提名了 4 个男性和 1 个女性作为关系客体，那性别构成 IQV 就是 $\frac{1-[(0.8)^2+(0.2)^2]}{(2-1)/2}=0.64$；当某人提名 3 个男性和 2 个女性时，那得到的性别构成 IQV 就是 0.96。因此，后一自我中心网络展现了更大的性别多样性。所有关系主体的 IQV 均值构成了一个样本定性多样性的分类变量指数。利用这一方法，Marsden(1987)报告说美国人自我中心网络的性别 IQV 是 0.68，而种族 IQV 只有 0.05。很明显，美国人的核心讨论圈在性别构成上要比在种族差异上均衡得多。

全幅网络测量 在自我中心网络研究中，一个中心个体通过列举与他/她有特定关系的客体来构建其网络。除非从一个高密度社区抽样——其中关系主体和客体之间都联系紧密(White & Watkins, 2000)，否则某一个主体的关系客体极少与另一主体的关系客体有关系。因此，之前部分中讨论的密度测量是建立在互无联系的知情人所提出的关系客体的特征基础之上的。相反，在全幅网络研究中，研究人员设定一个网络边界，只关注在此范围内的行为人之间的联系。例如，Krackhardt(1987)的研究关注某一组织内 21 位高级技术经理人。每一位都被要求确认是否从其他 20 位经理人那里获得过建议。利用全幅网络设计获得的数据，只由所选出的这些行为人的联结构成。通常，选定范围之外的行为人的联结是不会考虑的。因此，把任何一种自我中心网络密度测量法应用到全幅网络中只得到单个的值。N 个行为人的随机抽样，产生 N 个自我中心网络密度值，因为每一个行为人的自我中心网络都不同于其他行为人。相反，全幅网络有一个群体的单个密度值，不论总体 N 是多少。尽管在分析层次上有这种差异，但对于自我中心网络和全幅网络而言，这 4 种密度测量的计算公式都是相同的。

除密度之外，可广泛适用于全幅网络的其他关系测量方法还包括节点度、行程、轨迹、路径、回路、组成部分、测地距、直径、分界点、桥、树、同构子图等等。本书无法做到对所有这些方法进行详尽阐述，相关资料可见 Wasserman 和

Faust(1994)。这里,我们讨论两个基本主题:节点度和测地线。这两种测量方法同样都可以用于全幅网络和自我中心网络。为方便讲解,我们以代表不同的网络数据类型的 4 个图标为例。

节点度表示为 $d(N_i)$,是行为人 i(也就是节点)的所有关系数量,其中度指线的条数。在一个二进制无向图中,如图 4.3 所示,每一个行为人的节点度为 $d(N_A)=4$, $d(N_B)=d(N_C)=d(N_D)=d(N_E)=3$。该全幅网络的平均节点度,由所有行为人的节点度相加,再除以行为人总数,表示节点间的联结程度。对图 4.3 来说,平均节点度是 $\frac{4+3+3+3+3}{5}=3.2$。

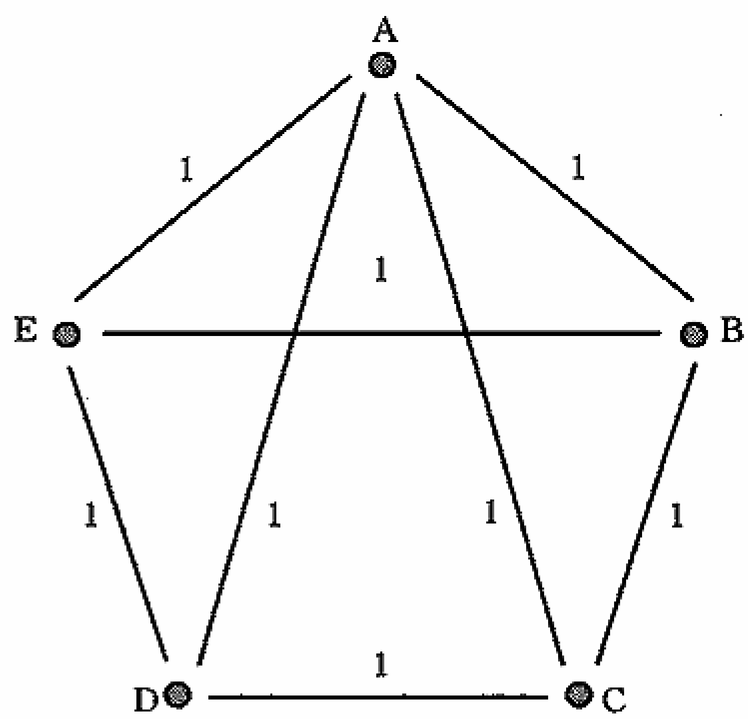


图 4.3 二分无向图

然而对于有向图,研究者必须区分节点入度和节点出度。节点入度是一个行为人接受的从其他行为人方向所来的线的数量;节点出度是从一个行为人节点向其他行为人所发出的线的数量。图 4.4是一个朋友提名网络,其中 A 提名 B和 E 作为朋友,但并没有其他人将 A 作为朋友。因此,A 的节点入度为 0,但节点出度为 2。虽然一个人的节点入度和出度会有差别,但全幅网络的平均节点入度和出度总是相等的,因为指向所有行为人的线条总数等于从所有行为人处发出的线条总数。在图 4.4 中,行为人 A、B、C、D、E 的节点入度分别为 0、2、1、1、3,但它们的节点出度为 2、1、1、2、1。每一套度数总和都是 7,因此网

络平均节点度数为 $\frac{7}{5}=1.4$。一些学者认为，一般意义上节点出度表示扩展程度，而节点入度代表接纳程度或者受欢迎程度(Wasserman & Faust, 1994: 126)。根据这种有趣的观点，A 是一个开朗型的行为人——会选择他人作为朋友，但是他是一个在网络中最不受欢迎的人。相反，E 是最受欢迎的，但相对来说他是个内向型的人。

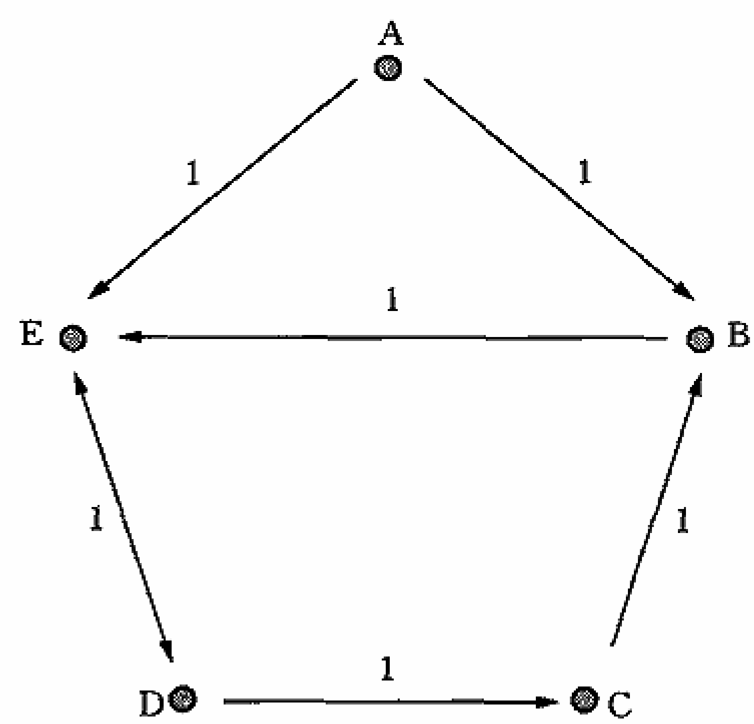


图 4.4　二分有向图

对于计算赋值图的节点度数，通过网络文献回顾可以得到些许启示。其中一个简化的解决办法，就是将赋值图中所有的非零值替换为 1，然后采用二进制图的处理办法。但是，这一方法忽略了赋值图中有关联结力度的丰富信息。在此，我们推荐一个概括的办法，用于计算赋值的无向图的节点度数。这一方法将一个行为人所有线条的赋值加起来。然后，该赋值的无向图的平均节点度数就是所有赋值节点度数的总和除以行为人的总数。图 4.5 中 A、B、C、D、E 的节点度数分别为 5、13、7、5、12。因此该图的平均节点度数是 $\frac{5+13+7+5+12}{5}=8.4$。我们承认，这一求和法在下述两种相反的情形之下，会产生较高的节点度数：行为人线条值低，但数量多；行为人线条数量少，但赋值高。例如，有一朋友关系网络，其中的线条赋值表示朋友关系节点的力度。

第一种情形下，一个人有许多关系浅的朋友；第二种情形下，一个人只有少许关系非常密切的朋友。概括方法为两种节点度数都提供了同样的高值，但是并不能将二者区分开来。

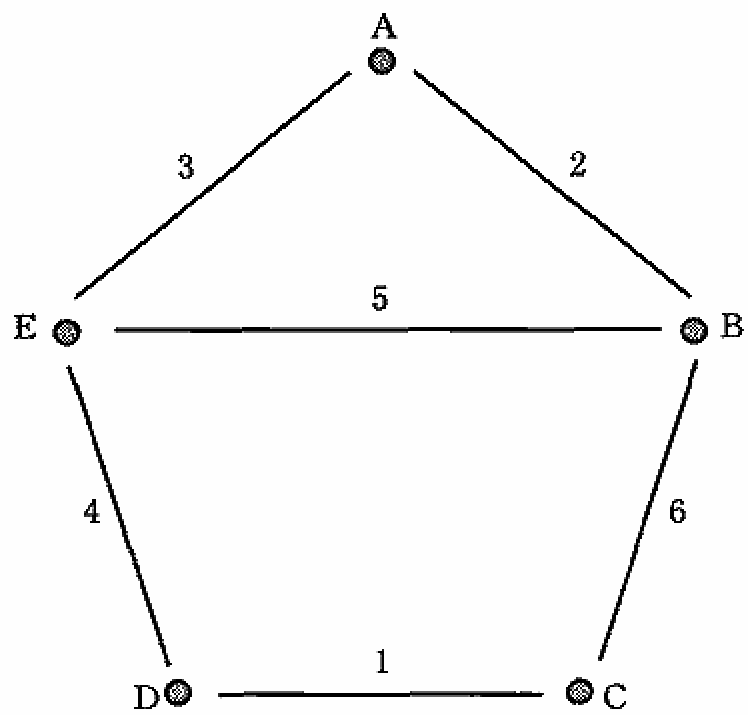


图 4.5 赋值无向图

解决这个难题的一个办法，我们称之为均值法，它将赋值的线条数量考虑进来，用某一行为人线条的赋值总数除以线条总数。因此，均值法计算与行为人直接联系的线条的平均赋值。这一解决办法可以区分前述段落中所论述的两种情形。

假定行为人 X 有 5 个关系浅的朋友，每个朋友关系赋值为 1；行为人 Y 只有 2 个朋友，但每个朋友关系都高赋值，分别为 2 和 3。求和法得到同样的节点度数 5。而均值法则各产生一个平均数，行为人 X 的节点平均数为 $\frac{1+1+1+1+1}{5}=1$，而 Y 的平均数为 $\frac{2+3}{2}=2.5$。Y 的平均节点度高于 X 表明 Y 比 X 有着更为紧密的朋友关系。

对赋值有向图的节点度数测量在使用节点入度或节点出度时，必须考虑关系的方向性。下表给出了利用求和法和均值法获得有 5 个关系节点的图 4.6 中节点入度和节点出度的结果。

	节点度	A	B	C	D	E
求和法	节点入度	0	8	1	4	12
	节点出度	5	5	6	5	4
均值法	节点入度	0	4	1	4	4
	节点出度	2.5	5	6	2.5	4

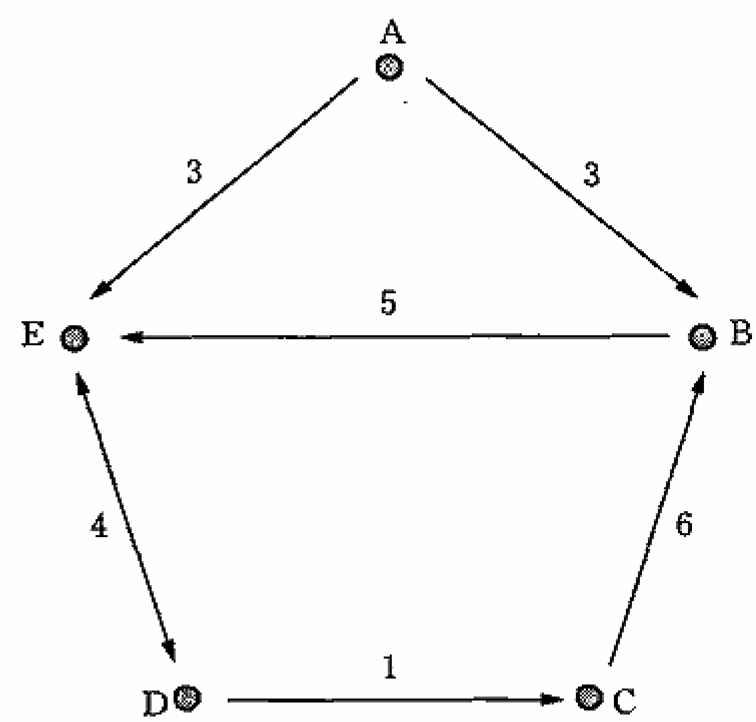


图 4.6 赋值有向图

对两种方法进行比较，可以发现几点有趣的现象。使用均值法，节点 B、D、E 有相同的节点入度，表明这三者同样受欢迎。相反，求和法则发现三个节点入度成较陡的层次区别，其中 E 最受欢迎，B 次之，D 位于最末。如图 4.6 所示，求和法显示 E 得到程度不等的最多提名，因此将 E 视为是最受欢迎的点。均值法强调，虽然 D 只得到一次提名，但是它有较高的赋值，因此 D 的排名与 E 和 B 一样。根据他们不同的研究目的，网络研究者可以选择不同的测量方法来测量有向赋值图的节点度，但必须谨慎解释他们的结果，解释的时候必须和求这些值的方式一致。

测地距是两个行为人之间最短路径的长度。它的计算因全幅网络的四种类型不同而差别较大。让我们从最简单的例子开始：图 4.3 的二进制无向图。节点 A 和 B 有许多条路径相连（例如，AB，ACB，AEB，ADCB，ACDEB，

AEDCB)，路径长度从 1 至 4。最短路径(AB)的长度为 1，因此 A 和 B 之间的测地距是 1。图 4.3 包括 10 对关系 ($C_N^2 = \frac{5!}{2! \times 3!} = 10$)。这 10 对关系间的测地距，依两点之间是否有直接连接而距离不同。例如，AB 的测地距是 1，而 CE 的测地距是 2(通过 CDE，CBE 或 CAE)。

含有 N 个行为人的有向图的双向连接数是同等规模的无向图的两倍，因为在有向图中，每对行为人有两个有方向的连接，而在无向图中则只有一对连接。例如，图 4.4 中 AB 之间有两个有向关系：AB 和 BA。由 A 连至 B，有两个路径：AB 和 AEDCB(但 AEB 并不是一条路径)。最短路径是 AB，因此 AB 间的测地距是 1。反过来，B 不能连接至 A，因此 BA 之间的测地距不明确。有向图的测地距分析，在网络研究中有许多实际的运用。其中一种情形，就是绘制一个行为人群体之间的交流图。假定图 4.4 表现了一个 5 人间的交流网络，A 可以直接向 B 发送信息，或者通过 E、D、C 向 B 间接地发送信息。但是通过中介传送信息容易延误或者歪曲，因此正如 AB 间的测地距所分析的那样，A 最有可能直接与 B 交流。相反，B 既不能直接也不能间接地向 A 发送信息。实际上，A 只能充当一个信息发送者，而不能作为一个信息接收者，因为没有人与 A 交流。这个例子生动地描述了信息的流动并非是互惠的。

当考虑每一条路径的赋值时，计算赋值有向图的测地距会面临一些困难。我们可以将赋值图简化成二进制图，然后运用二进制图处理程序，但因此也浪费了这些赋值所提供的丰富信息。Yang 和 Knoke(2001)提出了一个解决之道，利用路径长度和路径值来测量赋值图双方的测地距。这一方法通过一个合适的运算法则来实行(Yang & Hexmoor, 2004)。大体上，这个方法涵盖：(1)找出所有的连接路径；(2)找出每一条路径的最小值；(3)用最小值除以每一条路径的长度；(4)在第 3 步的结果中挑选出最高的值，这便是一个赋值图的测地距(也称为最佳连接，Yang & Knoke, 2001)。这个方法假定，长的路径最容易受到延迟和扭曲，而最短的路径对于双向交流来说等同于瓶颈路段。在图 4.5 中，连接 A 和 B 有 3 条路径，他们的平均路径值为：AB = 2，AEB = $\frac{3}{2}$ = 1.5，

AEDCB $= \frac{1}{4} = 0.25$。选择最高的平均路径值，那 A 和 B 之间的测地距为 2，因此最佳路径就是 AB。

Yang 和 Knoke(2001)的方法可以运用于赋值有向图中两点间的测地距的计算，不过关系的方向必须被考虑其中。图 4.6 连接 A 到 B 的两条路径是 AB 和 AEDCB，其中 AB 平均路径值是 2，AEDCB 的平均路径值是 0.25。选择最大的路径值，那 AB 间的测地距就是 2。然而，连接 BA 间的测地距是无法确定的，因为 B 无法通过任何路径到达 A。

节点度和测地距是社会网络分析的重要测量方法。节点度比较容易计算，并且常能为实际运用提供大量信息。节点度一般表示一个行为人参与网络行为的程度。例如，朋友关系网络中的节点度低，表示该行为人朋友较少。对全幅网络而言，平均节点度表示行为的综合水平。例如，一个社区住户间串门的平均节点度较低，意味着邻居间活动少。节点度也是节点度中心度测量的一个基础，这个在下文中将要讨论到。两节点间的测地距是节点之间关系亲密度非常重要的标准。由于每对行为人间都有几条路径相连，因此测地距就是每两节点间的最短路径，也是基于距离的综合法的基础，例如多维标度和多层聚类分析，这些在本节的后面将会进行讨论。

中心度和声望

图论在社会网络分析中的一个主要作用是，在个人分析层次和群体分析层次上，找出其中重要的或突出的行为人。中心度和声望的概念及测量通过综合所有点之间的结构关系，试图将图论关于全幅网络中行为人知名度的理念进行量化。对知名度的分析，将考量社会计量学上的选择与被选择，有时也包括间接关系的结构。个人行为人的知名度反映了他/她在其他行为人当中较为明显的可见性。群体层次上的中心度和声望指数，评估所有行为人的知名度在一个网络中的分散程度或不平衡程度。在本部分中，我们讨论两个有关知名度测量标准的大致分类：(1)中心度，无论是作为关系的发起者还是接受者，当一个行

为人知名度较高时，他/她对多种关系的参与程度较高；(2)声望，表示知名行为人虽然只发起较少的关系，但是接受很多直接关系。知名度测量的选择，取决于概念化的运用和可得网络数据的类型。它们的计算要求运用前面几节讨论的矩阵数学概念。现在我们从最简单的例子开始，先讨论无向图，接着再讨论有向图。由于绝大多数知名度指标只要求二进制联系测量，因此我们的讨论就限于这类数据，但是我们鼓励网络研究学者致力于建立赋值图的知名度测量方法。

最广泛使用的中心度测量标准是程度、亲密性和中间性(Freeman, 1977, 1979)。这些测量在运用于无向关系和有向关系时有所区别，在运用于个体行为人和群体或全幅网络层次时也有所区别。下面的计算公式利用了 Wasserman 和 Faust(1994)的表示方法。

程度中心度　在无向二进制图中，行为人程度中心度测量社会网络中一个节点与所有其他节点相联系的程度。对于一个拥有 g 个行为人的无向图，行为人 i 的程度中心度是 i 与其他 $g-1$ 个其他行为人的直接联系总数，用矩阵表示如下：

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij} \ (i \neq j) \qquad [4.1]$$

其中 $C_D(N_i)$ 表示节点 i 的程度中心度，$\sum_{j=1}^{g} x_{ij}$ 用于计算节点 i 与其他 $g-1$ 个 j 节点($i \neq j$，排除 i 与自身的联系；也就是说，主对角线的值可以忽略)之间的直接联系的数量。$C_D(N_i)$ 的计算就是简单地将行为人 i 在社会关系矩阵中对应的行或列所在的单元格值加总(因为无向关系构成一个对称性数据矩阵，因此行和列相同的单元格的值相同)。

如此测量的行为人程度中心度，不仅反映了每个节点与其他节点的关联性，而且也视网络规模(g)而定。也就是说，网络规模越大，程度中心度的最大可能值就越高。因此，一项具体的程度中心度值可表示一个小网络内行为人的关系紧密，也可以表示在一个大网络内行为人只与不多的几个人联系。为了消除网络规模变化对程度中心度的影响，Wasserman 和 Faust(1994:179)提出了

一个标准化的测量公式：

$$C'_D(N_i)=\frac{C_D(N_i)}{g-1} \qquad [4.2]$$

这个标准化行为人程度中心度测量，使用行为人 i 的程度中心度值除以与其他 $g-1$ 个行为人最大可能连接数，得到与行为人 i 有直接联系的网络成员的比例。

这个比例从 0.0 到 1.0，0.0 表示与任何行为人都没有联系（例如一个孤点），1.0表示与每一个人都有直接联系。标准化的行为人程度中心度测量行为人在诸多关系中的参与程度。得到高分的行为人是网络中最显眼的参与者。尤其是，标准化程度中心度值越接近 1.00，那么行为人在关系网络中的参与度越高。假如研究者更看重行为的纯粹数量而不是无向关系的具体来源和目标的话，那么他们很容易将中心度这一概念运用于信息网络的获取、控制和居间行为的分析（Knoke & Burt，1983）。

与行为人的程度中心度不同，组程度中心度测量一个社会网络中的行为人与其他行为人在个人程度中心度方面的不同程度。组程度中心度非常接近描述统计学的离差测量，例如标准差，表示围绕中心趋势值变化或扩散的量。Freeman（1979）提出了组程度中心度的通用测量方法：

$$C_A=\frac{\sum_{i=1}^{g}[C_A(N^*)-C_A(N_i)]}{\max\sum_{i=1}^{g}[C_A(N^*)-C_A(N_i)]} \qquad [4.3]$$

其中，$C_A(N^*)$表示网络中所观察到的行为人程度中心度的最大值，而 $C_A(N_i)$ 是 $g-1$ 个其他行为人的程度中心度。因此，分子是行为人最大中心度与其他所有人的中心度之差的总和。分母是这些差值之和的最大理论可能值。

按照 Freeman 的通用测量方法，Wasserman 和 Faust（1994：180）在他们的组程度中心度指数中提出了有关分母的处理方法：

$$C_D=\frac{\sum_{i=1}^{g}[C_D(N^*)-C_D(N_i)]}{(g-1)(g-2)} \qquad [4.4]$$

分子表示最大中心度的行为人与其他节点之间的程度中心度所观测到的差值之和，分母测量差值的最大可能总和。这个值可能发生在星形图中，其中一个节点与其他所有节点相连，但其他所有节点都只与第一个节点相连。因此第一个节点具有最高的可能程度中心度为 $(g-1)$，然而其他每个节点的程度中心度都等于 1，两者之间中心度的差值就是 $(g-1)-1=g-2$。因为这个差值在该图中发生 $(g-1)$ 次，因此分母值为 $(g-1)(g-2)$。组程度中心度指数值在 0.0 至 1.0 之间。当社会网络的程度中心度均匀分布时，即每一个点都具有相同的程度中心度时，$\sum_{i=1}^{g}[C_D(N^*)-C_D(N_i)]$ 等于 0；因此，组程度中心度为 0。在另一极端情形下，当一个节点具有最高可能的中心度 $(g-1)$，而其他所有节点的程度中心度为 1 时，表示最大的不均匀分布。分子等于分母，组程度中心度指数为 1.0。因此，组程度中心度越接近 1.0，社会网络中节点的程度中心度就越不均匀或越呈等级分化状。

亲密中心度 行为人亲密中心度是用于反映在一个社会网络中某一节点与其他节点之间的接近程度(Sabidussi, 1966)。亲密和距离指一个行为人能多快与他人联系上，例如通过直接交流或者极少的中间环节。行为人的亲密中心度是它与其他点之间测地距的函数(如前所述，测地距是连接两节点间的最短路径长度)。行为人亲密中心度指数为行为人 i 与 $g-1$ 个其他行为人之间测地距总和的倒数：

$$C_C(N_i)=\frac{1}{\left[\sum_{j=1}^{g}d(N_i, N_j)\right]}(i\neq j) \qquad [4.5]$$

这个测量值不可能为 0.0，因为在数学中 0 是不可能做分母的。因此，孤点(从节点 i 至该节点的距离是不可测的)亲密中心度值是不可能测量的，该指数只能用于可形成连接图的节点。对于实证社会网络分析来说，这一限制要求每一个行为人在网络连接图中至少要与其他一个成员存在连接关系。极低的行为人亲密中心度值，或中心行为人与其他人之间的高距离值，一般是因为该行为人处于一个较大的网络之中，或该行为人处于一个小网络中，但与该网络中的其

他行为人有相对较长的距离值。

当某一行为人与所有其他人关系亲密时，即在网络中与每个人都有直接联系时，$C_C = 1/(g-1)$，该值随网络规模大小发生变化。

为了控制网络规模便于不同网络间行为人的比较，Beauchamp(1965)建议，通过乘以 $(g-1)$ 来标准化行为人亲密中心度指数：

$$C'_C(N_i) = (g-1)(C_C(N_i)) \qquad [4.6]$$

为了比较和对比亲密中心度与标准化亲密中心度指数，我们参考两个例子。图 4.7 展示了一个 3 节点网络结构图，其中节点 A 直接与节点 B 相连，测地距为 1；间接与节点 C 连接，测地距为 2。因此，图 4.7 中节点 A 的亲密中心度是 $1/(1+2) = 1/3$。

图 4.8 展示了一个 4 节点网络结构图，其中所有点之间都有直接连接。行为人 A 与 B、C、D 都有直接联系，测地距均为 1。因此，行为人 A 的亲密中心度值为 $1/(1+1+1) = 1/3$。虽然图 4.8 中的行为人 A 与其他关系人的直接联系要比图 4.7 中的行为人 A 更多，但二者的亲密度值是一样的，因为第二个网络更大。为便于比较，标准化亲密中心度将网络规模考虑其中，对这两种情形进行了区分。图 4.7 的行为人 A 的标准化亲密中心度值为 $(3-1)(1/3) = 0.67$，而图 4.8 的行为人 A 的标准化值更高一点，等于 $(4-1)(1/3) = 1.0$。行为人的亲密中心度值越高，那么它与其他行为人的关系越紧密，从某种意义上来说，就是该行为人可以通过较短的测地距通达其他所有的节点。

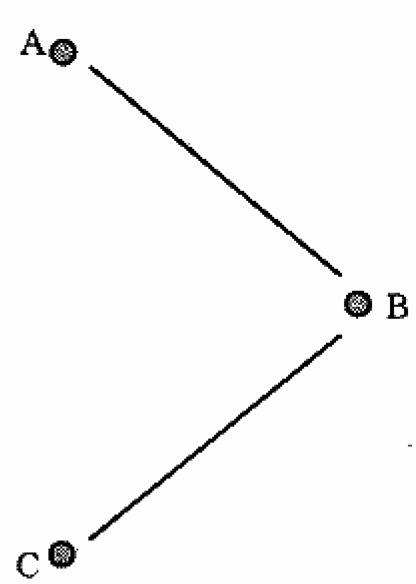


图 4.7　3 节点无向网络

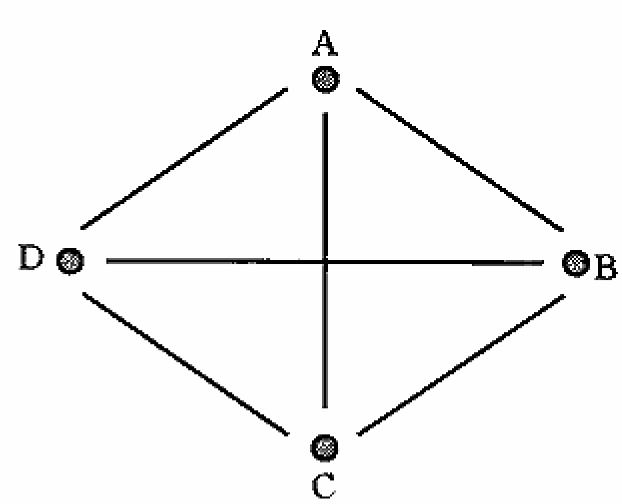


图 4.8　4 节点完全连接网络

与组程度中心度相似，组亲密中心度也是一种离差测量，表示在一个网络内亲密中心度的等级分层。具体来讲，组亲密性中心度测量一个给定网络内行为人各自亲密中心度之间的差异程度。根据 Freeman(1979)的文献，组亲密中心度指数的计算公式如下：

$$C_C = \frac{\sum_{i=1}^{g}[C'_C(N^*) - C'_C(N_i)]}{[(g-2)(g-1)]/(2g-3)} \qquad [4.7]$$

其中 $C'_C(N^*)$表示网络中可观测到的行为人亲密中心度的最大值，而 $C'_C(N_i)$是 $g-1$ 个其他行为人的亲密中心度。当一个网络中行为人的亲密中心度分布极其不均时，组亲密中心度达到最大值，为 1.0。此时，某一行为人具有最大的亲密中心度，而其他所有人只有最低的亲密中心度。相反，如果每一个行为人都有相同的亲密中心度，那组亲密中心度等于 0。

中间性中心度　中心度的中间性这一概念关注其他行为人如何控制或者调整并不直接连接的两节点间的关系。行为人中间性中心度测算其他行为人在何种程度上依赖网络中两节点间的测地距(即最短距离)。中间性中心度是一个非常重要的在一个网络控制信息交流或资源流动的指标。假设行为人 j 必须通过行为人 i 才能与行为人 k 沟通，因此行为人 i 负责或控制行为人 j 与 k 之间所传递信息的内容及时间。行为人 i 在众多对偶关系点间的测地距路径中出现的次数越多，行为人 i 控制网络交流的可能性就越高。

为了量化行为人 i 的中间性中心度，Freeman(1977)提出下列处理步骤。首先，g_{jk}是点 j 和 k 间测地距的路径数量，$g_{jk}(N_i)$是点 j 和 k 之间所有经过点 i 的测地距路径的数量。然后，将 $g_{jk}(N_i)$除以 g_{jk}，便得到连接点 j 和 k 的所有经过点 i 的测地距路径的比例。将所有不包括点 i 的对点加起来，便是测算节点 i 在其他网络成员间测地距路径的位置程度。下面的公式反映了这个逻辑：

$$C_B(N_i) = \sum_{j<k} \frac{g_{jk}(N_i)}{g_{jk}} \qquad [4.8]$$

当 $g-1$ 个节点间所有对点的测地路径都不经过点 i 时，该中间性中心度值为 0.0。假设每一对点只有一条测地路径，当节点 i 经由所有对点的每一条测

地路径时，该值即达到最大值 $(g-1)(g-2)/2$ $\left(\text{除节点 } i \text{ 之外，} g-1 \text{ 个节点之间的所有测地距路径数量是 } C_{g-1}^{2}=\frac{(g-1)!}{2!(g-1-2)!}=\frac{(g-1)!}{2!(g-3)!}=\frac{(g-1)(g-2)}{2}\right)$。然而，假如每对点有不止一条测地距（最短路径）时，节点 i 的中间性中心度的最大可能值将超过 $\frac{(g-1)(g-2)}{2}$，具体值依每对点间的测地距路径的数量而定。

Wasserman 和 Faust(1994:190)通过 $C_B(n_i)$ 除以 $\frac{(g-1)(g-2)}{2}$ 的理论极大值来得到标准化行为人中间性中心度 $C'_B(N_i)$（当然，还是假定每对节点间只有一条测地距路径），其数学表达式为：

$$C'_B(N_i)=\frac{C_B(N_i)\times 2}{(g-1)(g-2)} \qquad [4.9]$$

当初始中间性中心度是 0 时，标准化行为人中间性中心度值也是 0.0；当其他 $g-1$ 个节点中每对节点间的测地距路径都含节点 i 时，标准化行为人中间性中心度是 1.0。因此，该值越接近 1.0，节点 i 就越能控制或调解网络关系。

如同组程度和亲密中心度测量一样，组中间性中心度测量该值在网络中不同行为人间的变化程度。按照 Freeman(1979)的基本测算方法，Wasserman 和 Faust(1994)提出了组中间性中心度指数：

$$C_B=\frac{\sum_{i=1}^{g}[C_B(N^*)-C_B(N_i)]}{((g-1)^2(g-2))/2} \qquad [4.10]$$

分子是最高值行为人和每个其他行为人的中间性中心度值的差异总和。分母表示网络中所有节点的中间性中心度理论上的最大可能值。行为人中间性中心度在 $\frac{(g-1)(g-2)}{2}$ 时获得理论最大值。在群体层次，如果某一主要节点居间促成了所有其他的测地线，那这个最大值至少发生 $g-1$ 次。因此，对于一个含有 g 个行为人的网络，其理论最大中间性中心度值是 $\frac{(g-1)^2(g-2)}{2}$。

(这里我们再次强调，这个计算假定每一对点只有一条测地距路径。如果两点之间存在多条测地距路径，那么行为人中间性中心度的最大可能值将大于 $\frac{(g-1)(g-2)}{2}$，这将导致理论最大可能值的相应变化。)

当所有的测地线都经由某一主要节点时，组间中心度为 1.0。相反，如果在一网络中每一个节点都有相同的中间性中心度，那么分子为 0，组层次中心度也为 0.0。因此，中间性中心度越接近 1.0，网络内中间性中心度越是不均衡分布。

声望 许多情况下，社会网络数据收集包括指导如何在调查中确定网络关系的发起人或发送人和终结者或接收者。在这样的有向图中，仅仅参与或融入某一关系，要比区分这一关系的发起者和接收者次要得多。例如，在一个工作场所的报告关系网络中，低阶位的员工通常向他们的管理监督者报告他们的工作行为和绩效工资绩点，而高层雇员很少或根本就不用对他们的下属报告工作。在一个朋友关系网络中，一个热情地将许多人视为好朋友的人，可能没有人将他视为好友。

我们将声望定义为一个网络中的一个社会行为人接收他人所发出的关系或被他人视为关系客体的程度。这种发送者—接收者或源头—目标的区别，极大地强调了对资源控制上的不平等，以及伴随这种不平等而来的权威和服从(Knoke & Burt, 1983:199)。

根据这个定义，声望可以简单地通过计算某一个行为人接收其他行为人某种特定关系的有向连接数量来测量。例如，Laumann 和 Knoke(1987:163)询问美国能源和健康政策领域的知情人士，请他们确认“对制定国家能源/健康政策具有特别的影响力和重要性”的机构名单。每一个提名都得到平等对待，没有将其来源进行区别对待，这样所获提名便反映了一个行为人在系统内的受欢迎程度。Wasserman 和 Faust(1994:202)指出，行为人程度声望可以根据有向图的节点入度来计算：

$$P_D(N_i) = \sum_{j=1}^{g} x_{ji} \ (j \neq i) \qquad [4.11]$$

因为这一测量加总由 g 个节点所组成的非对称社会矩阵中的第 i 列的值，

因此它的最小值是 0.0，最大值是 $g-1$。一个标准化的行为人程度声望，是控制了网络规模的大小，可以对不同网络进行声望比较：

$$P'_D(N_i)=\frac{\sum_{j=1}^{g}x_{ji}}{g-1}(j\neq i) \qquad [4.12]$$

当所有其他行为人都提名行为人 i 时，标准声望值为 1.0；当无人提名行为人 i 时，标准声望值为 0.0。因此，这一比例越接近 1.0，行为人的声望越高。

上述段落中所介绍的中心度、组中心度和声望的基本测量方法，都是基于二进制数据的。除了 Freeman、Borgatti 和 White(1991)讨论的赋值图中的中心度问题之外，对赋值图的相关测量研究极少。很显然，对于赋值图的中心度和声望测量问题，需要更多的研究。

自我中心网络中心度 中心度和声望的分析局限于全幅网络数据，直到 Marsden(2002)将中心度的测量延伸至自我中心网络数据。他讨论了一个单一关系的二进制对称数据的中心度测量。在使用调查对象进行的自我中心网络研究中，关系主体提名 g 个与他有特定关系的客体。每一个关系主体 i 产生一个 $(g+1)\times(g+1)$ 的数据矩阵 A_i，它的行和列都由关系客体 1 至 g 加上主体 i 构成。因为在定义中，主体与所有其他的客体具有直接联系，因此第 i 行中的每一个单元是 $x_{i,j}=1$，其中 $1\leqslant j\leqslant g$，而且 $i\neq j$。

在一个全幅网络中，行为人程度中心度是行为人与他人的直接联系（参见方程 4.1）。在含有 g 个关系客体的自我中心网络中，关系主体 i 的程度中心度是行为人程度中心度的最大可能值 $g-1$。关系主体 i 的标准化程度中心度总是 $\frac{g-1}{g-1}=1.0$。全幅网络中的行为人亲密中心度，是行为人 i 与其他行为人之间测地距总和的倒数（参见方程 4.5）。标准化亲密中心度控制了网络规模（参见方程 4.6）。因此，行为主体的亲密中心度和标准化亲密中心度分别为 $\frac{1}{g-1}$ 和 $\frac{g-1}{g-1}=1.0$。最后，全幅网络中的中间性中心度测量节点 i 经过所有对点间测地距的位置程度（参见方程 4.8）。由于行为主体 i 在其自身的自我中心网络中与所

有的行为客体直接连接，因此 i 可以作为每对并不直接相连的客体间的中介点。

Marsden(2002:410)宣称，在自我中心网络中，节点 i 的中间性中心度的测量不同于全幅网络。一方面，在自我中心网络中，节点 i 的中间性中心度会偏小。自我中心中间性测量不包括主体节点 i 的中间位置，这中间位置是任何一对经由超过三个路径长或更长的测地距的节点之间的位置，但是在全幅网络的中间性测量中则包括这种情形。另一方面，当两个客体节点的连接不但可以经由节点 i 连接，而且还经由节点 i 所处自我中心网络之外的其他节点时，自我中心网络中间性中心度会夸大主体节点 i 的中间性中心度。在这种情况下，自我中心网络中间性测量会忽略此网络之外的中间节点，而全幅网络中间性中心度测量则会包括这种中间节点。

关于这些观点的描述，请参看图 4.9。行为主体 i 的自我中心网络由在某种关系下提名的关系客体 A、B、C、D 组成。全幅网络也包括节点 M，它连接节点 i 的 2 个客体(A 和 D)，虽然 M 本身并不是节点 i 自我中心网络的一部分。节点 i 的 4 个关系客体构成 6 对关系(AB、AC、AD、BC、BD、CD)，其中每一对都有一条独特的测地距路径。由于只有 BC 是直接相连的，而其余每对关系的测地距路径都包含主体点 i，因此 i 的中间性中心度在自我中心网络中是5/6。然而节点 M 的出现，使得节点 i 的自我中心网络中间性中心度和全幅网络中间性中心度之间产生了差别。一方面，节点 i 自我中心网络中间性测量偏小，这是由于忽略了节点 i 在两条测地距(对点 MB 和 MC)中的媒介位置。另一方面，由于忽略了 AD 间的测地距通过点 M 的情形(这与通过节点 i 的 AD 间测地距形成了竞争)，节点 i 的自我中心网络中间性值又会偏大。

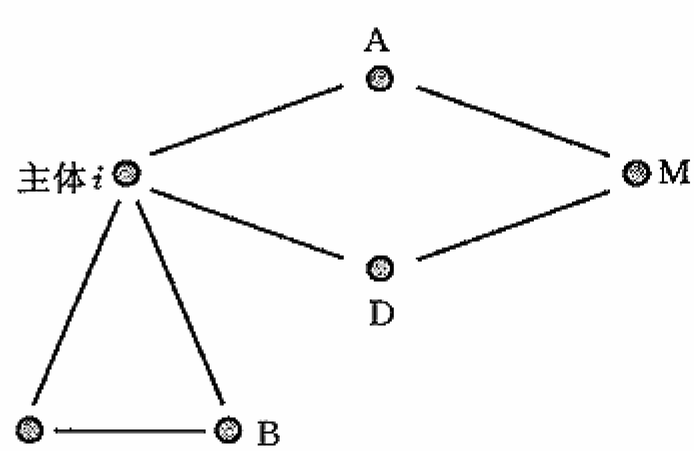


图 4.9 自我中心网络的中间性中心度

尽管在自我中心网络和全幅网络中间性的测量中存在这些差异，但Marsden(2002)通过分析17个实证网络发现这两种测量方法相差不大。因此，当全幅网络中相关数据难以获得时，自我中心网络中间性中心度似乎可以作为全幅网络的行为人中间性中心度的可靠替代。Marsden(2002)认为，自我中心网络和全幅网络中间性结果的差异也许是由于两种网络数据收集的不同造成的，而这是一个值得进一步系统性研究的重要课题。

小团体

内聚力与一个社会群体或封闭社会关系圈中的亲密程度的概念密切相关。内聚型亚群体由通过多种直接互惠的选择关系相连的行为人构成。他们之间能够共享信息，团结互助，一致行动。所有的亚群体成员间有多种直接联系，很少或几乎不包含外来者，倾向于构成一个在思想、认同和行为方面具有同质性的群体。内聚型亚群体的例子包括异教组织、恐怖活动小组、犯罪帮派、军事行动小组、运动团队、手工行业和工作团队等。"小团体"已经进入了日常语言之中，主要指学校、教会和俱乐部这样的高层群体。这个术语在社会网络分析中具有特定含义，源于诸如节点、线和路径长度这些图论要素。利用小团体的概念来分析研究群体结构，可以帮助研究人员更好地理解内聚力是如何通过提供建议和利益支持来帮助群体成员的(Dunbar & Spoor, 1995)，更好地理解对小团体的广泛信赖是如何限制网络联系的范围的(Blau, Ruan & Ardelt, 1991)。

社会学的概念，诸如组、群、圈子、帮派、派系、小团体等并没有严格的区别，可以互换(Borgatti, Everett & Shirey, 1990)。在对社会网络研究中大量关于亚群体的文献进行总结后，Wasserman 和 Faust(1994:251)分析出了关于内聚型亚群体的四个普遍特征：连接的相互性、亚群体成员间的可联络性、成员间联系的频率，以及与圈内成员间相对于与非圈内成员间联系的相对频率。这些特征是定义小团体和网络亚群体相关测量的基础。

小团体　一个小团体是含有3个或多个节点的最大完全子图，其中所有的节点都互相直接连接，网络中的其他点都没有与小团体中的每一个成员有直接

联系。因此，在一个小团体中，每对节点间的测地距都是 1.0。也就是说，每两个节点间都由一条直接路径相连。图 4.10 描述了一个含有 6 个节点的二进制无向性网络中的小团体。两个小团体 ABE 和 BCDE，每一个都符合必要条件。虽然 F 与 A 和 C 都有连接，但这 3 个点并没有构成一个小团体，因为 A 和 C 没有直接联系。4 个其他的小群体(EBC, EBD, BCD, BDE)也不是小团体，因为另外还有其他的点与这些小组别的所有三个点直接连接。例如，EBC 并不是一个小团体，因为点 D 与 E、B、C 直接相连。只有 BCDE 这个最大的完全子图符合小团体的条件。

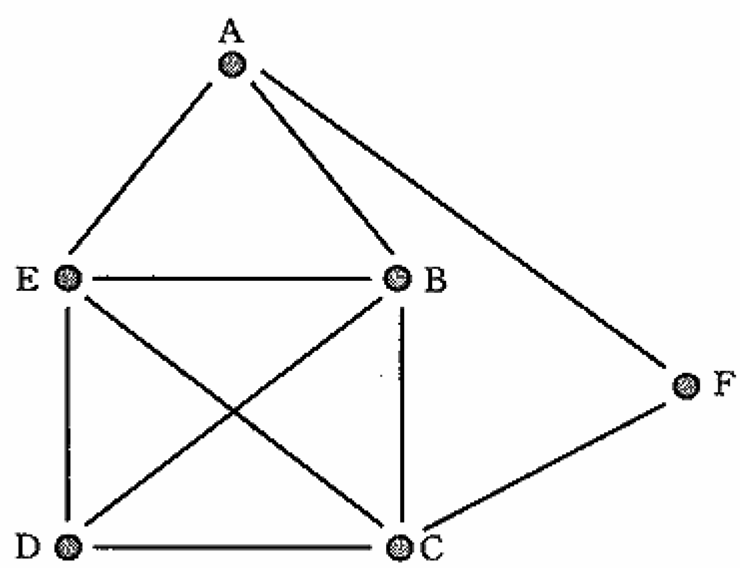


图 4.10 无向图中的小团体

作为识别内聚型亚群体的一个方法，小团体确认规则有着非常严格的条件。小团体概念有一个“吝啬鬼”的名声(Alba, 1973)，这是因为仅仅某一条线的缺失也许就会让一个子图不能成为一个小团体。因为小团体构成的高门槛，实证研究者极少在低密度的实际网络中发现大的小团体(Wasserman & Faust, 1994:256)。在实际网络数据中很少发现小团体的原因之一在于网络数据收集的设计。例如，固定名单法可能会限制受访者提名的关系客体的最大数量(如“说出你的三个最好的朋友”)。因此，小团体的规模不可能超过研究人员设定的上限。

小团体标准也严格地将内聚型亚群体内成员与群体外成员区别开来，因此忽略了更核心的和更边缘的行为人之间的等级层次。一旦小团体的边界划定了，在

小团体和非小团体成员间的区分就很难更为精细了。如此简单的一分为二法则常常是不够的,因为它不能反映更为重要而微妙的差别。一个图中小团体数量的决定因素之一是全幅网络的规模。小规模网络极少形成小团体,而大规模数据库常常产生大量的微小团体,而且它们还常常互相重叠。这一限制导致研究者集中在多重重叠的小团体,而不去搜寻一个单一的独立小团体。

n-小团体　虽然严格的小团体定义使得它拥有的信息量有限,但是它紧接着也催生了有关亚群体内聚力的测量,从而在一定程度上缓解了小团体构成方面的严格要求带来的压力。一个通常的方法就是基于节点度数,即通过某点的线条的数量(Doreian & Woodard, 1994; Seidman, 1983)。我们在第 3 章论述 *k* 中心概念时讨论过这个方法。这一方法将 *k* 中心定义为一个亚群体,条件是每一个群体成员至少与 *k* 个其他点直接连接。通过改变 *k* 值,研究者能够为定义一个亚群体的边界设定一个严格或宽松的标准。

第二个通常的方法是利用节点连接来修正小团体(Alba, 1973; Mokken, 1979)。*n*-小团体概念放宽了小团体本身的严格条件,即每对节点间的测地距值必须是 1.0。在 *n*-小团体中,所有对点间的最大测地距是路径长度 *n*;也就是说,没有节点离其他每个节点的距离超过 *n* 个连接。通过改变 *n* 的值,网络研究者可以区分具有更强凝聚力的亚群体(*n* 小)或者更弱凝聚力的亚群体(*n* 大)。例如,设定 *n* 值等于 2,即等于一个作为有 3 节点或多节点的子集的 2-小团体,其所有的测地距是 2 或更小;而网络的其他所有节点都与 2-小团体连接,其测地距是 3 或大于 3。在无向图(图 4.10)中,节点 F 在严格标准下不属于任何一个小团体。

然而,将测地距从 1 放宽至 2,即可得到一个包含节点 F 的 2-小团体(整个图为 ABCDEF),因为该节点(同时节点 A)连接 2-小团体的所有成员,而其中的测地距等于或小于 2。*n*-小团体标准具体化为:任何对点间的最大测地距不能超过 *n*。一般来说,*n* 的值越大,*n*-小团体的包容性也就越强,因此其成员间的凝聚力也就越弱,因为一些对象之间距离要达到 *n* 步。严格定义的小团体是 *n*-小团体的一个特例,其 *n* 值等于 1.0。

有向图中的小团体　有向图将关系的发送者和接收者区别开来。在有向

图中识别小团体需要注意对偶连接关系的互惠性。为拓展无向图中小团体的严格定义，我们将有向图小团体定义为由等于或大于3个节点构成的子图，其中所有的每一个节点都直接与其他节点连接，而且每对点之间都是对接的(即对偶连接)。为了描述有向图的小团体，图4.11展示了一个由6个节点组成的有向连接网络。在这个网络中存在3个3节点小团体，分别是ABE、BDE、BCD。由于从节点E到C的关系是非互惠的，因此EBCD不构成一个有向小团体。同样，A和C之间由于对偶关系的缺失也导致AFC不能成为一个有向小团体。

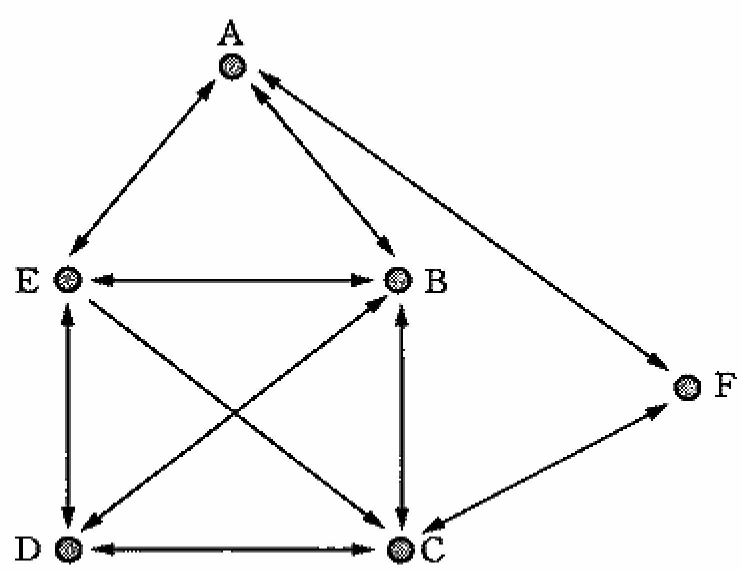


图 4.11 有向图中的小团体

从无向图 n-小团体的定义得到启发，研究人员提出，通过改变对偶连接的对点间的测地距值，可以放宽小团体定义上的严格要求。这一方法与前面的方法具有极大的相似性，除了更多的考虑有向连接问题。例如，两节点可以通过4种情形生成 n-连接：弱 n-连接，单向 n-连接，强 n-连接，双向 n-连接(Peay, 1980; Wasserman & Faust, 1994:275)。在最强形式的 n-连接——双向 n-连接——中，从 i 到 j 的最短路径和从 j 到 i 的最短路径必须通过同样的节点，不过是方向相反而已，而且其路径长度等于或小于 n。这四种情形的每一种都构成了一个有向 n-小团体，产生了4种类型的 n-小团体亚群体。在图4.11中，子图ABCDF是一个双向2-连接，因为它的10个对点通过测地距1或2都是互相可及的。然而，E不包括在内，因为它得不到C的反向连接，导致由E至C的测地距长要短于由C至E的两条测地距长(C-D-E和C-B-E)。

结构等价

社会科学家常常不仅对行为人的内聚力感兴趣，还对行为人的平等状态感兴趣。这种平等指在一个网络中两个或多个行为人与他人具有同样的或非常相似的关系。结构等价的行为人之间一般来说具有竞争型而非内聚型关系。譬如，两个大白菜种植者都将他们的产品卖给同一批蔬菜批发商，因此在出售大白菜时他们的关系就是结构等价，处于激烈竞争状态。结构等价的行为人完全可以互相替代。如果一个种植者退出蔬菜供应网络，那么他的位置很容易就被另外的结构等价的种植者替代，从而保持原来的网络不发生变化。为了从其他网络参与者那里获取适宜的反馈，社会网络中的完全可替代性常常导致激烈的竞争，众所周知的例子就是小学生为了获得老师的特别青睐而互相竞争。利用结构等价方法的网络研究学者往往对理解竞争关系有兴趣，而不是对群体内聚力有兴趣(Burt, 1992)。

与小团体的定义类似，结构等价的定义也是非常严格的。在有向二进制图中，如果两个行为人在与他人发送和接收关系时具有完全相同的连接模式，那么他们在那种具体关系上就是完全结构等价的。更详细地讲，对于网络中所有的节点 k(但不包括节点 i 和 j)，如果节点 i 发送一个连接给 k，当且仅当节点 j 也发送一个连接给 k；如果节点 i 从 k 接收一个连接，当且仅当节点 j 也从 k 接收一个连接，那么节点 i 和 j 就是结构等价的(Wasserman & Faust, 1994: 356)。对于多重网络来讲，如果两点结构等价，则上面讲的这个条件必须在 R 种关系的每一个当中都要精确把握。节点 i 和 j 之间有向连接的存在与否，对于决定它们是否结构等价是无关的。它们的结构等价仅仅是由它们与 $g-2$ 个其他网络点的关系模式来决定。

无向二进制图不区别关系的发送者和接收者。因此，将上述有向关系结构等价的定义沿用至无向图，如果对于网络中所有的其他行为人 k，i 与 k 具有连接，当且仅当点 j 也与 k 具有连接，那么 i 和 j 就完全是结构等价的。结构等价也可以用在赋值图中，其中利用排名等级而非二进制赋值来测量两节点之间的

连接。严格来讲,只有当赋值图的两个节点在与其他节点的每一个连接中都具有完全同样的值,这两个节点才是结构等价的。

对于实证网络分析的实际运用来说,完全结构等价性的上述定义通常太过严格。真实网络数据中极少有数据能够满足这样苛刻的标准。然而,当一些节点与其他节点的关系模式高度相似,尽管不是完全相同,但我们也可以认为这些节点是大致结构等价的。为了获得这种近似性,研究人员利用关系相似性的测量,而不采用严格的、非此即彼的完全结构等价的标准要求。两个节点各自与其他所有节点的连接越相似,那这两个节点就越接近结构等价。

基于网络关系相似性的两节点间的结构等价,只要求他们与其他节点或有或无的关系模型高度相似即可。假定存在一个二进制有向图,两个结构等价的行为人,在其社会矩阵对应的行和列中的条目非常雷同。根据这一原则,Burt (1978)提出将欧氏距离作为测量行为人 i 和 j 结构等价的一种方法:

$$d_{ij} = \sqrt{\sum_{k=1}^{g}[(x_{ik}-x_{jk})^2+(x_{ki}-x_{kj})^2]}(i \neq j \neq k) \qquad [4.13]$$

其中 d_{ij} 是行为人 i 和 j 之间的欧氏距离,$x's$ 是社会矩阵(第一个下标表示行,第二个下标表示列)的值(对于二进制关系来说就是 1 或 0)。由于 d_{ij} 是两个差值的平方和的正平方根,因此每个 $d_{ij} \geqslant 0$。假如行为人 i 和 j 与其他行为人之间拥有非常一致的连接,那么所有的差值 $(x_{ik}-x_{jk})$ 和 $(x_{ki}-x_{kj})$ 都是 0;因此这两个完全结构等价行为人的欧氏距离 $d_{ij}=0$。但是在绝大多数实际例子中,观察到的 d_{ij} 值要大于 0;因此,行为人 i 和 j 在某种变动的程度上关系类似。欧氏距离与行为人的相似性相反,因此也与结构等价相反:d_{ij} 值越大,行为人 i 和 j 的结构等价度就越小。也就是说,欧氏距离实际上测量的是一对行为人之间的异质性。

为了阐述如何计算对点之间的欧氏距离,图 4.12 和表 4.3 分别描述了一个 5-节点网络结构的有向图和矩阵式。图 4.12 表明,行为人 A 和 B 是完全结构等价,因为他们都与行为人 C 和 D 有直接联系,但都与行为人 E 没有联系。相反,行为人 D 和 E 就不是结构等价了,因为尽管他们都发送了一个连接给行为

人 C,但是行为人 D 接收了行为人 A 和 B 的连接,而行为人 E 却没有得到任何连接。利用表 4.3 的二进制值,可以计算行为人 A 和 B 的欧氏距离:

$$d_{AB}=\sqrt{[(x_{AC}-x_{BC})^2+(x_{CA}-x_{CB})^2]+[(x_{AD}-x_{BD})^2+(x_{DA}-x_{DB})^2]+[(x_{AE}-x_{BE})^2+(x_{EA}-x_{EB})^2]}$$

$$d_{12}=\sqrt{[(1-1)^2+(0-0)^2]+[(1-1)^2+(0-0)^2]+[(0-0)^2+(0-0)^2]} \qquad [4.14]$$

其中 $d_{12}=0$,表明行为人 A 和 B 是完全结构等价。你能否算出点 D 和 E 之间的欧氏距离是 $\sqrt{2}=1.41$?

由于无向二进制图不区分关系的发送者和接收者,因此它的欧氏距离计算就简单得多:

$$d_{ij}=\sqrt{\sum_{k=1}^{g}(x_{ik}-x_{jk})^2}\,(i\neq j\neq k) \qquad [4.15]$$

当网络呈现多重关系时,欧氏距离就得计算所有 R 种关系的对点差值的平方和:

$$d_{ij}=\sqrt{\sum_{r=1}^{R}\sum_{k=1}^{g}[(x_{ikr}-x_{jkr})^2+(x_{kir}-x_{kjr})^2]}\,(i\neq j\neq k) \qquad [4.16]$$

与通过欧氏距离来计算异质性不同的是,皮尔逊相关系数 r_{ij} 直接测量关系的相似性,其中对点关联值越高,表明结构等价度越大。我们将在本节块模型部分讨论相关性问题。

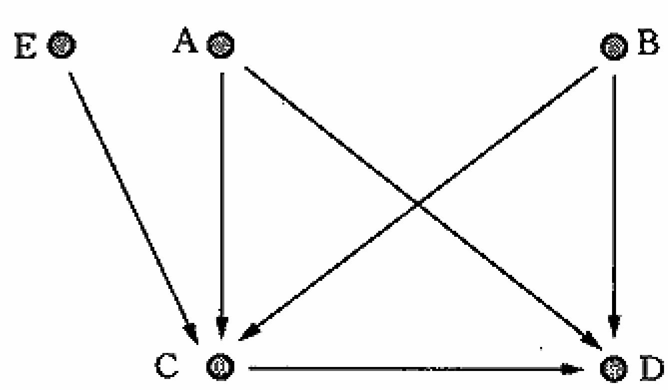


图 4.12　5-节点网络的结构等价有向图

表 4.3 5-节点网络的结构等价矩阵

	A	B	C	D	E
A	—	0	1	1	0
B	0	—	1	1	0
C	0	0	—	0	0
D	0	0	1	—	0
E	0	0	1	0	—

视觉演示

网络图像在社会网络研究中普遍用于展现和传递网络结构的亮点。社会网络分析在视觉演示的发展过程中经历了三个明显的阶段(Freeman, 2000, 2005)。起始阶段大概在 20 世纪 30 年代,当时 Jacob Moreno 发明了一种手绘的即兴社会网络图,用来描绘诸如学生等行为人之间的关系(Moreno, 1934)。这种自由式的方法很快就让位于通过标准计算程式来标示点和线的图。可视化初始阶段的一个基本原则就是所示空间应该代表行为连接的模式,即在数据矩阵中社会关系最近的对点,在可视图像中也能看得出来是距离最近的(Freeman, 2005)。然而,在二维或三维的视觉展示图中,精确标注多位行为人之间关系的亲近度和距离,通常是不可能完成的任务。因此,研究者最终采用系统简化、维度减少但依然能够反映原始数据模型的办法。可视化第二阶段始于大约 20 世纪 60 年代,用大型电脑主机和软件自动生成图表。尤其需要指出的是,网络分析不断利用这部分将要讲到的分层聚类法和多维标度法。最近的阶段始于 20 世纪 90 代中期,随着因特网、高速计算机网络和浏览器,以及个人电脑普及化时代的到来,关系数据库的大型可视化展示,包括纵向网络变化的动画展示,就有了新的机会(例如 Brandes, Kenis, Raab, Schneider & Wagner, 1999; de Nooy, Mrvar & Batagelj, 2005)。

因为篇幅有限,我们就集中讨论分层聚类和多维标度这两个展示网络数据的基本理念。可视化展示对于用社会网络数据来探索发现内聚型亚群体以及

揭示它们内部的关联性是非常有用的。在此过程中，可通过将众多行为人概括为数量较少的共同占据位置，对复杂的网络结构进行简化。结构等价方法计算和分析对点异质性（欧氏距离）或相似性（关联指数）的矩阵模型，来确认由行为人与其他人相同或相近关系模型所构建的亚群体所占据的网络位置。聚类和多维度标示法运用于结构等价矩阵，来确认共同占据位置，以二维或三维图的形式来展示结果。

聚类 分层聚类分析，或者简称为聚类分析，将行为人分成亚群体（共同占据位置）。这些亚群体成员是标准的或近似的结构等价。每一个行为人开始都被当成一个单独的聚类，然后这些单独的聚类依次联合起来，直到合并成为一个独立的聚类。一个树图，也被称为树形图，形象地展示了这个合并聚类的层次顺序。

聚类演绎法 一般处理 $g \times g$ 阶方阵的欧氏距离（d_{ij}）或关联指数（r_{ij}），其中 g 是具体关系中的行为人数量。聚类分析的一项比较研究发现，这两种方法都得到非常相似的结果，虽然使用相关性概念的聚类分析会相对容易解释（Aldenderfer & Blashfield, 1984:24—28）。多重网络关系（R）可以同时被聚类，通过构建一个 $g \times g \times R$ 维的数据阵列，在进行聚类分析前，计算所有 R 矩阵中的每对点间的欧氏距离或相关性。虽然分层聚类可用于二进制数据和赋值数据，但为简化的缘故，我们主要对欧氏距离的二进制对称矩阵进行讨论。

在计算对点的欧氏距离矩阵之后，聚类分析开始根据一个阈值 α 对行为人进行合并。这个阈值作为分析者的一个上限，让他们决定在一个结构等价的特定水平上，哪一个行为人属于同一个关联性占据位置（聚类）。仅当 $d_{ij} \leqslant \alpha$ 时，行为人 i 和 j 共同占据一个位置。在一个聚类中的行为人互相之间的社会距离要小于他们与其他聚类中行为人的距离，即聚类内的行为人更具有结构等价性。通过聚类演绎法，不断将 α 的值放宽（即 α 值变大），并运用于整合行为人的过程，直到整个网络合并成一个单一的全涵盖聚类。尽管分层聚类法获得一些非交叉聚类，但聚类还是嵌套的；也就是说，较小的聚类依次包含在 α 值高的较大聚类中（即，结构等价度越低，相似性越小，那么聚类内距离就越大）。最终，

研究人员必须决定哪一个层面的块(即哪一个 α 值)能够为网络中结构等价位置数量提供最好的替代。

研究者可从形成聚类的三个基本标准中进行选择:单个联结,平均联结,或完全联结。给定 α 值的水平,当两个关系最为亲近的行为人距离小于 α 时,单个联结标准将两个聚类合并成一个聚类。在完全联结标准下,当每一对行为人的距离小于 α 时,两个聚类进行合并。平均联结标准是一个折中的选择,要求两个合并的聚类中两类行为人的平均距离小于 α。实证研究发现,每一种选择都存在有利和不利的一面,研究者应当根据具体的情况来调整所选的聚类分析方法(Aldenderfer & Blashfield, 1984:53—62)。完全联结聚类倾向于得到数量众多的同质型和紧凑型的聚类,而产生“链”的可能性较低。“链”在这里指依次增加一个行为人即可形成一个单一的较大的聚类(Burgin, 1995; Wasserman & Faust, 1994:381)。

前面的内容主要讨论了小团体、聚类和结构等价等有关的概念和方法。这些网络分析方法普遍用于揭示社会结构的基本要素,诸如社会内聚力(小团体和聚类)或结构性竞争的程度(结构等价性)。因为这些方法探查社会结构的中间层次,因此他们承上启下,将社会网络的微观模式与社会结构的全景网络连接起来。下面我们讨论多维标度,这是揭示网络行为人结构的另一个非常重要的方法。

多维标度 多维标度(MDS)是将数据结构进行可视化处理的第二个基本方法。MDS 的主要目的是,发现那些能够反映网络行为人之间相似性和差异性(距离)的有意义的维度。如同聚类分析所述,MDS 的典型要素是一个欧氏距离为(d_{ij})或关联指数为(r_{ij})的 $g \times g$ 阶矩阵,其中 g 是具体关系中行为人的数量。MDS 的可展现结果是一张图,或称为社会关系图,其中较短距离(较大相似性)的关系行为人比较远距离(较大差异性)的关系行为人在空间上要离得近一点。虽然 MDS 图能够展示多个维度,但大多数分析一般都只是表达二至三个维度。

在 MDS 图中所表示的行为人之间的距离与输入矩阵图中的相似性或差异性值相关但不完全相同。这个距离反映的是通过 MDS 程序所估算到的这些数

据中对点间的距离。应力这一概念测量所观测到的矩阵和所计算到的矩阵二者之间所有点的差异性(Kruskal & Wish, 1978:23—30):

$$Stress = \sqrt{\frac{\sum\sum(f(x_{ij}) - d_{ij})^2}{Scale}} \quad [4.17]$$

其中,$f(x_{ij})$是输入值的一个非计量性的、单一的函数(Kruskal & Wish, 1978: 29),d_{ij}是呈现在图中行为人i和j之间的欧氏距离。Scale是一个应力值在0.0至1.0之间的测量指示因子。如果一个MDS图将一组输入数据转化成标准的图值,那么对于所有的i和j来说$f(x_{ij} = d_{ij})$,应力将等于0。所以,应力值越低,MDS图的空间就越能够代表所观测到的网络行为人的社会距离。基于这个原因,应力概念也被称为MDS拟合度的测量法,其中应力值低于0.1被认为是拟合得非常好,应力值在0.1至0.2之间被认为是拟合度适中,应力值大于0.2就拟合较差(Kruskal & Wish, 1978:52; Slez & Martin, 2007)。

MDS曾经被运用于各种研究。例如,Slez和Martin(2007)利用MDS对各州一段时期内随时间变化投票的相似性进行绘图比较。Whit, Kim和Glick(2005)对多伦多地区的50个种族群体绘制了社会空间位置图。MDS的最大优势,是允许他们同时分析所有的种族群体,而传统的分隔研究方法则将每一个群体与所有其他群体进行分离对比。

为了描述聚类和MDS这两种方法是如何能够一起联手揭示结构等价网络位置,我们可以看图4.13和图4.14。图4.13分层聚类树图和图4.14 MDS图都是基于欧氏距离,而该欧氏距离是从图4.12的二进制网络数据计算出来的。聚类分析使用完全联结选择:当所有行为人之间的距离小于α值时(这些值显示在图4.13图的水平轴上),聚类合并在一起。例如,行为人B和E间的欧氏距离(α)是0.00(标准的结构等价),因此树图首先在欧氏距离等于0.00这一阶段将B和E合并。有两个聚类出现在欧氏距离等于1.00这一阶段:一个聚类合并了行为人A和C,另一个聚类将D与B—E这一现存的聚类合在一起。在欧氏距离$\alpha = 1.73$时,这两个聚类合并了,表明所有对点的欧氏距离都小于1.73。最后,在欧氏距离$\alpha = 200$时,行为人F与A—B—C—D—E聚类合并,表明F

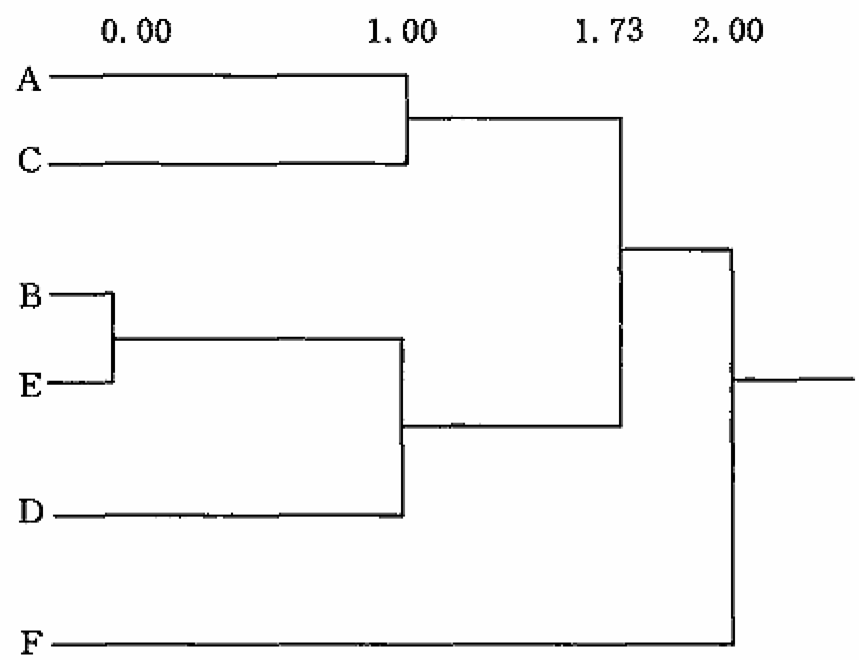


图 4.13　图 4.12 的网络分层树形图

和所有其他节点之间的欧氏距离在 1.73 至 2.00 之间。

图 4.14 的 MDS 图中，行为人亚群体周边有邻接线，这个位置正好是在聚类树图中 $\alpha=1.00$ 所确认的三个聚类所占的位置。

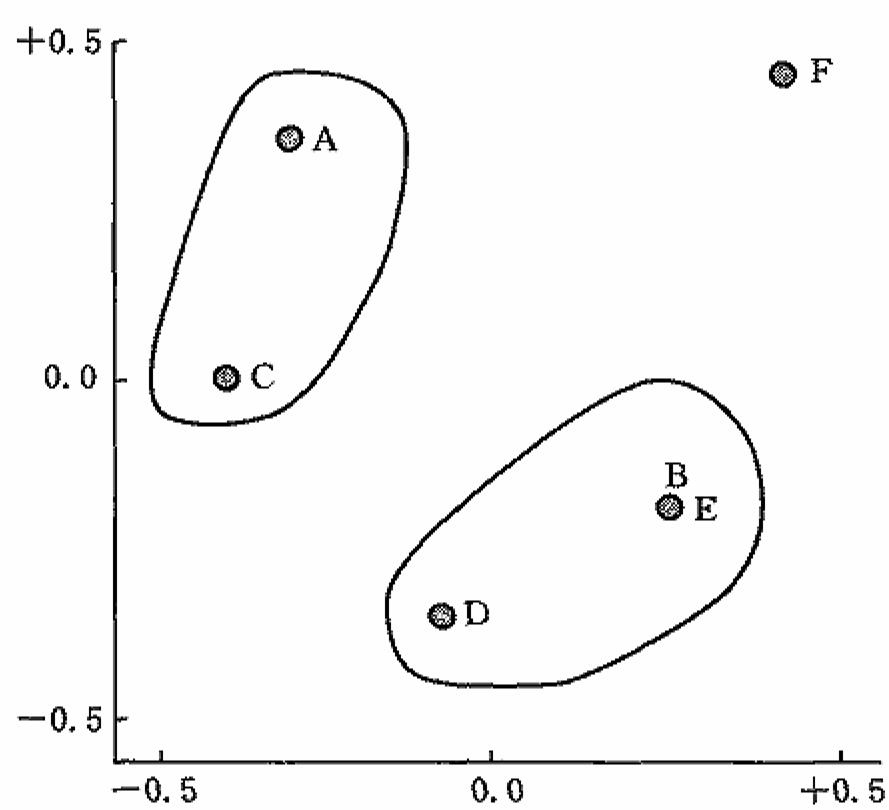


图 4.14　图 4.12 的网络多维标度

行为人 B 和 E 在图中正好占据着同样的点，因为他们是完全结构等价。虽然对于确认聚类来说，α 值的选择在一定意义上是随意的，但我们仍然建议选择

一个值,它能够平衡网络行为人之间的内聚力和分裂性。例如,如果我们已经知道欧氏距离为 0.00,那么我们将有 5 个聚类:B 和 E 构成一个聚类,另外 4 个行为人占据单件聚类的位置。为方便比较,现假设我们选择 $\alpha = 2.00$,那么结果只有 2 个聚类:(A—B—C—D—E)和(F)。第一个选择标准,由于其亲密关系设置上的高门槛,产生了大量的分裂性;而第二个值,则使得门槛降低,得到了较大的内聚力。然而,没有一个方法能够像中间值 $\alpha = 100$ 时提供的信息,即揭示联合占据位置之间的结构关系,为这个网络的分离性与内聚力提供一个更平衡的视角。

MDS 的一个延伸运用是加权多维标度(WMDS)。它在传统的 MDS 方法中增加了加权因素,用于代表矩阵间变化的信息。因此,WMDS 允许多重矩阵被进行系统性地区分,从而对距离模型进行概括。例如,假如每一个矩阵都对应一个不同的个人,那么 WMDS 利用权重就可以对这些个人思考和想象他们的关系进行不同的刻画。基于这个原因,WMDS 也被称为个体差异标度(INDSCAL)。虽然我们不能在这里更详尽地讲述 WMDS,但我们鼓励有兴趣的读者向 Schiffman, Reynolds 和 Young(1981:55—85)咨询;还有 Han(2003),他对此方法有最新的运用。

块模型

块模型是一个矩阵代数方法,将网络行为人分类成联合占据的、结构等价性的位置。块模型方法开始由 Harrision White 和他的助手研发(Boorman & White, 1976; Schwartz, 1977; White et al., 1976)。自此开创性工作之后,研究者已经将块模型方法成功地运用到各种问题研究中,从跨组织网络(Knoke & Rogers, 1979)到新技术的扩散(Anderson & Jay, 1985),再到世界体系中的城市角色(Alderson & Beckfield, 2004)。方法研究者也加深了块模型搜索和划分法则的研究(Nowichi & Snijders, 2001; Winship & Mandel, 1983; Wu, 1983)。因为篇幅有限,我们在这里不继续讨论这个大的话题,在本节集中讲述块模型的基本问题及其应用和结果的阐释。

块模型是 g 个行为人社会网络矩阵的一个部分，它将一个或多个关系网络，分成两个或多个被称为块(B)的亚群体或位置。块这一术语指结构等价行为人的一个亚方阵，其中的行为人与其他块中的行为人具有相似的(如果并非完全相同的话)关系。建立块模型是一个数据简化技术，通过重新对行为人进行分组，提出浓缩的综合信息，它将对网络数据进行系统性搜索，从中找出关系模型。结果就是重新排列置换过的密度和图像矩阵，表示每一种类关系中块内部和块之间的联结方式。块模型可以被用于单个或多重的关系，有向或无向的联结，二进制或赋值的图。我们的块模型例子是基于 UCINET 的 CONCOR 运算法则，它能够同时处理二进制数据和赋值图。我们这里只讲述二进制数据矩阵，并推荐读者阅读 Dorian, Batagelj 和 Ferligoj (2005:347—360)的相关文献，以进一步了解二进制数据和赋值图的块模型分析。

可以利用理论原则事先构建一个块模型，例如将不同管理部门的雇员分类成不同的块。然而，该方法的最普遍运用是对关系数据的经验模式进行调查搜寻。建立块模型常常使用 CONDOR(Convergence of iterated Correlations)，这一方法在多个社会网络分析程序中都可以找到(Schwartz, 1977)。CONDOR 可以通过关联对行、对列，或者同时通过关联矩阵的两个向量进行运算。为教学之便利，我们这里仅关联列项。首先，CONDOR 计算二进制 $g\times g$ 阶社会网络矩阵图中每一对列项的皮尔逊相关系数(对于块模型 R 多重关系来说，独立的 $g\times g$ 矩阵是堆积起来的，即安排成为一个单独的二维数据阵列，其目的是输入给 CONDOR)。这些计算不包括每对点间的直接联结，因为它们的结构等价性仅仅依赖于对点与其他 $g-2$ 个网络行为人的联结关系。假如行为人 i 和 j 与其他行为人有着几乎完全相同的联系，那么他们的相关系数 r_{ij} 将等于 1.0。相反，如果两个行为人有着几乎完全相反的联系模式，那么他们的相关系数等于－1.0。实际的相关系数几乎总是在这两个极端值之间。从 CONDOR 开始得到的结果是一个有关每对行为人的 $g\times g$ 阶对称的关联矩阵，表示每对点结构等价性的程度。第二步，其后的所有迭代步骤，都重复着前一步关联矩阵中的对列关联过程。在一定程度上，矩阵每一格的相关

系数都趋向于 1.0 或－1.0，直到达到这两个点时迭代中止。

接下来，CONDOR 将最后的关联性矩阵排列成两个同质的块。社会网络矩阵的排列包括同时对行和列的重新安排，以将这些占据同样块位置的行为人一起排在相邻的位置。CONDOR 对社会关系网络的初始分区和排列总是得到两个分矩阵，它们在行为人的数量上不一定相等，在同一块内每对行为人的所有关联指数都等于 1.0，但是两个块之间的关联性却是－1.0。重复这一过程，CONDOR 能够将初始得到的两个块中的每一个都续分成两个，以此类推。网络研究者必须决定在何处停止这一分化过程，因此也就决定了最终所得到的块的数量。

块模型分析导致两个主要形式的结果：密度矩阵和相应的图像矩阵。密度矩阵是一个 $b\times b$ 阶矩阵，它的单元格的值是块内和块间的密度（密度是一个比例，其计算是由重新排列所得子矩阵中所观察的联结数量除以可能的所有联结数量）。图像矩阵也是一个 $b\times b$ 阶矩阵，通过将密度矩阵中的每一单元格的密度值编码为 0 或 1 转化而来。下列两个可替代标准可以用来决定图像值：(1)行为人之间没有联结的格（0-块）编码为 0，行为人至少有 1 个联结的格（1-块）编码为 1；(2)研究者选择一个密度临界值 α，那么在这个临界值之下的所有密度编码为 0，所有等于或高于 α 值的编码为 1。上面第一个选项是一个不现实的标准，因为实际密度为 0.0 的情形很少发生，除非一个网络中的关系很少，从而几乎所有图像矩阵中的每个单元格都有 1 构成。第二个选择是最经常碰到的情形，它利用 α 密度临界值对图像值进行二项分类。研究者都选择满矩阵的密度值作为临界值。然而，因为选择 α 值不可避免地要受到研究人员的判断问题，因此一个选定的特定值容易受到独断性方面的批评。为了回应此类批评，研究人员应当尽量在理论和实践上提出合理理由，而不能仅仅诉诸利用之方便（Scott，1991：136）。

作为块模型运用的实际范例，我们分析了美国 1988 年里根政府执掌行政权力时期国家劳工政策领域中的政策交流网络（Knoke et al.，1996）。这个政策领域的核心由 24 家最有声望和影响力的政治组织组成，由知情人士从 117 个该领域的组织中选出（参见表 4.4，组织名称首字母缩写）。图 4.15 是得到确认的

有关劳工政策事项的交流网络的有向图(对于一个存在的有向联结,一对组织必须认同所发出的信息也得到了接收)。在这个值(0.29)之内的高密度联结,使得这个图中任何简单的结构都难以辨认。对应的交流关系矩阵,在表 4.4 中被转化成一个同时利用行(发送者)和列(接收者)的关联矩阵。这些关联测量每对组织与其他 22 家核心领域成员的交流模式的相似性。之后,对这个关联性矩阵进行 CONDOR 分析,得到表 4.6 这样一个含有 4 个模块的区分、密度和图像矩阵。

为了决定图像中的 1-模块,选定的密度临界值(α)等于或高于0.33。最后,在图 4.16中关联性矩阵的 MDS 分析,利用围绕 4 个模块成员的邻接线,展示了核心组织之间的距离。

在社会关系图中,联合占据的块的构成及它们的定位,反映了 20 世纪后期美国的劳工和资本之间的基本政治裂度。在图 4.16 的右下角,模块 1 是一个由 4 个工会组织组成的支持劳工派,包括 AFL-CIO①,代表老年人利益的集团(AARP)和参议院劳工委员会的民主党派。较小的模块 2 包括 3 个利益集团和 ETA(联邦劳工部的一个下属机构)。模块 3 站在支持商业的立场,由里根白宫政府、劳工部、三个商业组织,还有卡车司机工会(它不是 AFL-CIO 成员)。模块 4 也主要是站在支持商业的立场,包括其他的 3 个国会委员会(2 个共和党群体和 1 个众议院民主党派)和 2 个主要的商业组织(商业圆桌会议和美国商务部)。根据表 4.5 的图像矩阵,每 1 个块包含着内部政策交流的高密度,但是在 12 个块间密度中只有 3 个块存在经常的联结:支持劳工派的块 1 发送政策信息给块 4 的国会部门,而国会部门所在的块 4 和支持商业的块 3 也互相交流信息。虽然块模型和 MDS 分析反映了它们在理念上的清晰区分,但一些组织(例如国会众议院民主党派人士和卡车司机工会)出现在未曾料想的位置上,意味着政策交流在朋友之间和敌对方之间都会发生,这也许可以缓和美国国家劳工政策领域政治对立的紧张程度。

---

① 劳联—产联,美国最大的工会联盟。——译者注

表 4.4　美国劳工政策领域核心交流网络的二进制矩阵图

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	AFL	0	1	1	1	1	0	0	1	1	0	0	0	0	0	1	1	1	1	0	0	1	1	1	1
2	ASCM	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
3	TEAM	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	0	0	1	1	1	1	1	1	0
4	UAW	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
5	USW	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
6	ABC	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	1	0	1	0	1	1	1	1	0
7	AGC	0	0	1	0	0	1	0	1	1	0	1	0	0	0	0	1	0	1	0	1	1	1	1	1
8	BRT	0	0	0	0	0	0	1	0	1	1	1	0	0	0	1	1	1	1	0	1	0	0	1	0
9	CHAM	0	0	1	0	0	1	1	1	0	1	1	0	1	1	0	1	0	1	1	1	1	1	1	0
10	GM	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0
11	NAM	0	0	0	0	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	1	1	0	1	0
12	NEA	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	AARP	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
14	ACLU	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1
15	NGA	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0	0	0	0
16	DOLS	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
17	ETA	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
18	OSHA	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
19	NLRB	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
20	WHO	1	0	1	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	1	1	0
21	HR	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0
22	HD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	SR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	SD	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0

AFL:美国劳工联合会	ASCM:美国州县市工人联合会	TEAM:卡车司机工会
UAW:汽车工人联合会	USW:钢铁工人联合会	ABC:建筑工人和承包商协会
AGC:通用承包商协会	BRT:商业圆桌会议	CHAM:美国商务部
GM:通用汽车公司	NAM:全美制造商协会	NEA:美国教育协会
AARP:美国退休人员协会	ACLU:美国民权同盟	NGA:全美州长协会
DOLS:劳工部	ETA:劳工部教育和培训管理局	OSHA:职业安全和健康管理局
NLRB:全国劳动关系委员会	WHO:白宫办公室	HR:众院共和党劳工委员会
HD:众院民主党劳工委员会	SR:参院共和党劳工委员会	SD:参院民主党劳工委员会

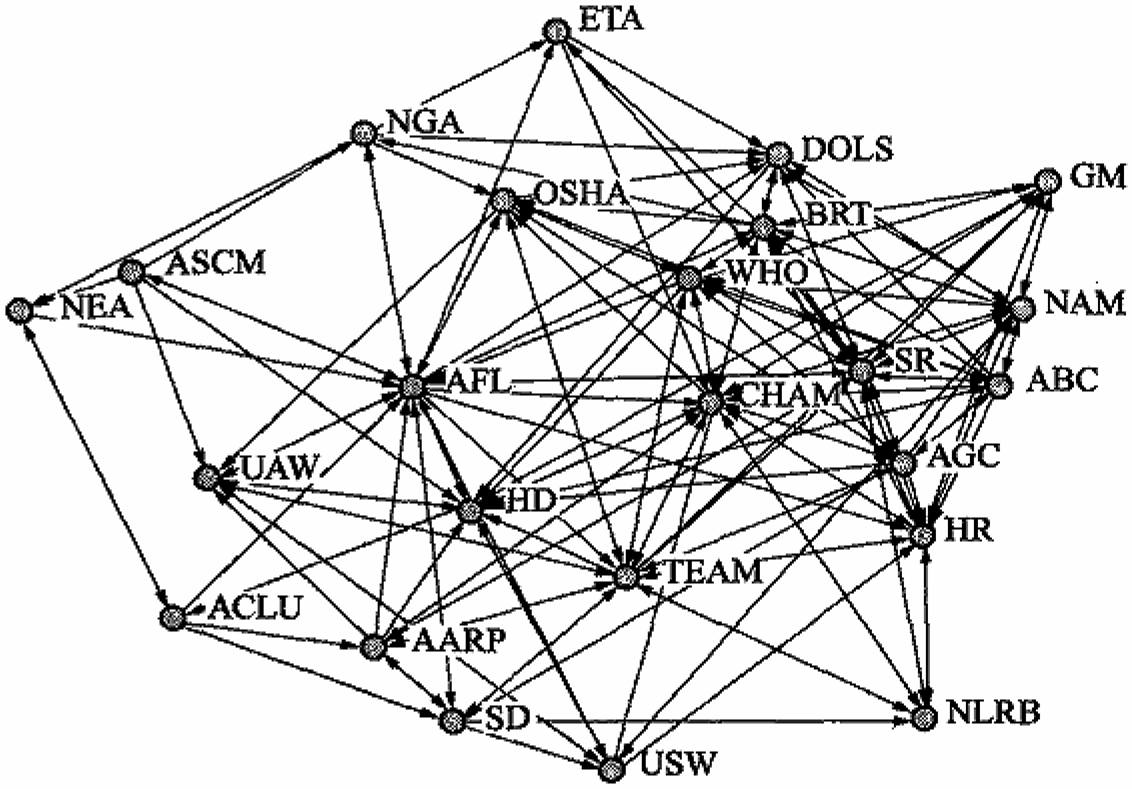


图 4.15 美国国家劳工政策领域关系有向图

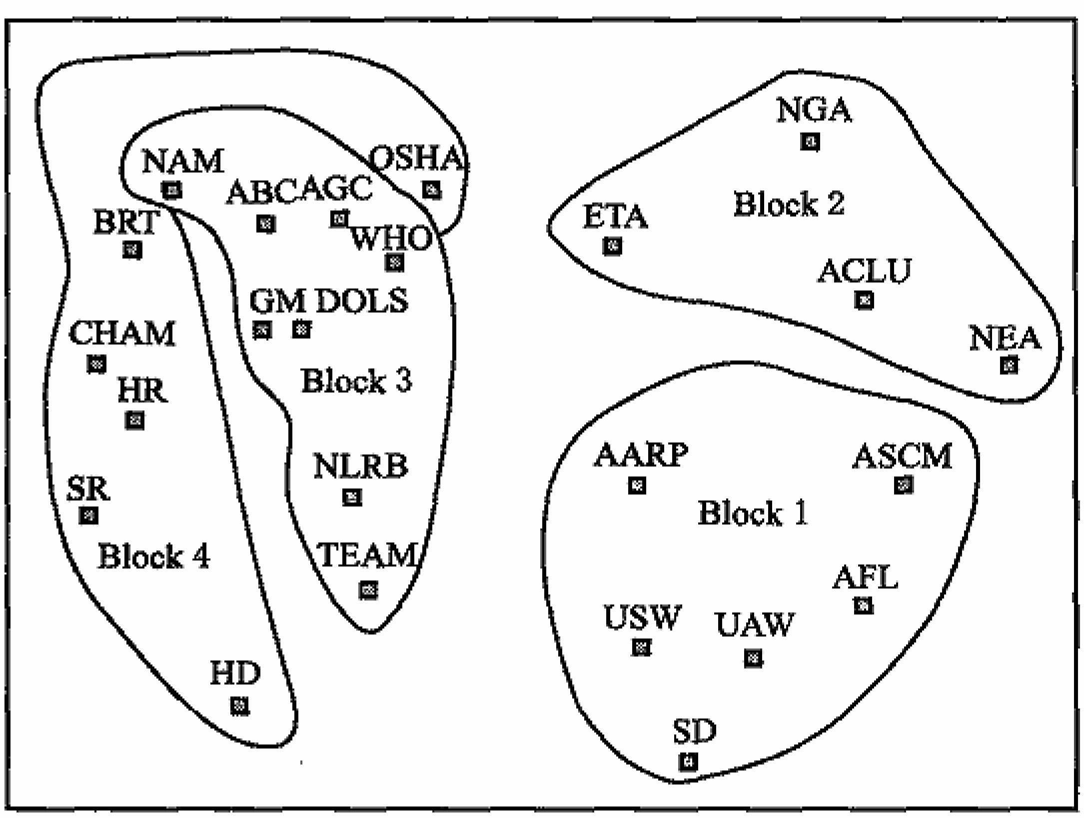


图 4.16 美国国家劳动政策领域关系的块状多维标度

表 4.5 美国劳工政策领域中的核心交流网络的分块阵、密度阵和图像阵

Blocked Matrix

	1	2	24	4	5	13	15	14	12	17	10	3	7	20	6	16	11	19	9	8	21	22	23	18
1 AFL		1	1	1	1		1			1		1				1			1	1	1	1	1	1
2 ASCM	1			1			1		1													1		
24 SD	1			1	1	1						1						1						
4 UAW	1		1		1							1										1		
5 USW	1			1																1	1	1		
13 AARP	1		1									1							1			1	1	
15 NGA	1								1	1				1		1								1
14 ACLU	1		1			1			1										1					
12 NEA	1							1																
17 ETA	1																1		1				1	
10 GM																	1		1	1	1	1	1	
3 TEAM				1		1							1	1				1	1		1	1	1	1
7 AGC			1									1		1	1	1	1		1	1	1	1	1	1
20 WHO	1										1	1	1		1				1	1		1	1	1
6 ABC													1	1		1	1		1	1	1	1	1	1
16 DOLS																			1	1	1	1	1	
11 NAM					1						1			1	1	1			1	1	1		1	
19 NLRB												1							1		1		1	
9 CHAM						1		1			1	1	1	1	1	1	1	1		1	1	1	1	1
8 BRT							1			1	1		1	1		1	1		1				1	1
21 HR											1	1				1	1	1	1	1			1	
22 HD																								
23 SR																								
18 OSHA				1								1	1		1	1								

Density Matrix

	1	2	3	4
Block 1	0.567	0.167	0.125	0.389
Block 2	0.250	0.333	0.094	0.167
Block 3	0.104	0.000	0.393	0.813
Block 4	0.056	0.125	0.458	0.367

Image Matrix

	1	2	3	4
Block 1	1	0	0	1
Block 2	0	1	0	0
Block 3	0	0	1	1
Block 4	0	0	1	1

第 5 章 | 网络分析高级方法

在这一章中，我们将讨论有关网络分析的高级方法。我们尤其要评论网络位置测量、logit 模型 p* 、隶属网络和点阵分析。

网络位置测量

在社会网络分析中，角色和位置是核心概念。利用代数和矩阵方法对互锁角色的分析，始于 S. F. Nadel(1957)关于社会结构的理论。在其理论中，他将社会结构视为位置，而该位置拥有与其他位置相关的独特的权利和责任。在第 4 章讨论过的结构等价，为确认社会网络角色和位置提供了一个基本的方法。然而，它要求结构等价的行为人与同样的其他行为人具有相同的连接模式，这一要求对于该方法的实际运用来说，太过苛刻。虽然放松关系相似性这一要求可以扩展该方法的运用，但是在行为人基于自身的关系模式联合占据位置方面，结构等价依然有着非常严格的定义。作为一个替代性选择，网络研究者创造出多个适用条件没有如此严格的方法来确认网络角色和位置(Borgatti & Everett，1992；Everett，1985；Everett，Boyd & Borgatti，1990；Faust，1988；Pattison，1988)。在适用条件严格的降序排列中，结构等价是最严格的，其次是自形和同构等价，再次是正规等价。本部分对后两种方法进行非技术性的一般回顾。虽然自形和同构等价经过一些调整能够被用于有向和赋值图(Wasserman & Faust，1994:461—502)，但为阐述的便利，我们还是在举例中利用了一个单个关系的无向二进制图。

同构和自形等价是两个关联性很强的概念，以至于一些研究者认为它们是

可以互相替代的(Borgatti & Everett, 1992)。但是,同构等价运用于两个图,而自形等价只能描述一个图中社会关系行为人的关系特征。如果一个图到另一个图的节点有着一对一对应关系,维持了所有节点的邻接关系(例如,同样的点入度和点出度),那么这两个图就展示了结构上的同质性。也就是说,如果在第一个图中两个节点相连,那么在第二个图中相应的两个点也必然以同样的方式相连(Borgatti & Everett, 1992:11)。每一个图自身是同构的,这就被称为自形形式,也就是一对一的对应关系映射到它自身。当且仅当两个节点与相应的其他位置(但不是同一个节点)连接时,这两个节点就是自形等价(即联合占据着同样的位置)。自形等价节点有相同的图论特征,比如中心度、自我密度、小团体规模(Borgatti & Everett, 1992)。

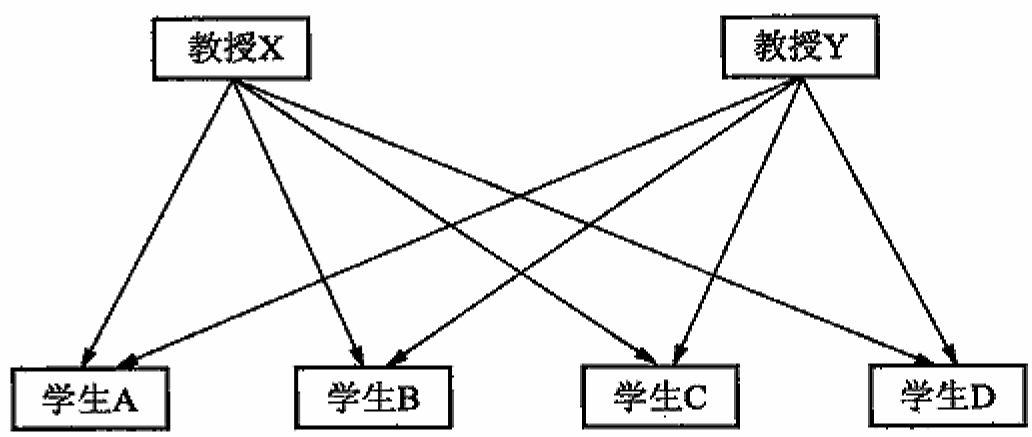
自形等价放宽了结构等价的要求,即在同一位置上的行为人与同样的其他行为人有着相同或非常相似的联结。相反,如果行为人与扮演该位置相同角色的不同行为人有着同样的联结,那么自形等价认同行为人共同占据着一个位置。举一个熟悉的例子:如果两个教授占据着结构等价的位置,那么他们两个必须教同样的一群学生,这实际上是不太可能的。但是,如果说要构成一个自形等价位置,这两个教授所教学生可以不同,但只需要所教学生数量相同即可。学生占据着第二个位置,由教授所教的人来定义。图 5.1 比较了这两种类型的等价,其中从教授到学生的有向连线表示"教学"关系。尽管这两个图显示有两个位置,但自形等价图更好地抓住了一个理念,即社会角色包含关系的一般模式。现在可以引述另外一个著名的例子:一夫一妻制婚姻。我们期待妻子的位置可由一群女性共同占据,她们与同一群男性有着非多妻制婚姻,但是她们每一个人都单独地与不同的丈夫形成配偶关系。

结构等价性行为人也是自形等价型的,但反之则不一定。如果行为人标签从图中移走,那么自形等价型的节点则不能区分。因此,如果图 5.1B 中的名字被点所替代,那么这两个子图就无法区分了。Borgatti 和 Everett(1992:16)如此总结这个区别:

"简要一点来说,我们可以认为,在结构等价方法中,网络或标示图可以代表一个群体的内在结构;因此,在该机构中的行为人的定位就代表了他/她在群

体中的位置。相反，在自形等价方法中，利益群体的结构并不是标示图本身；标示图被视为观测结构或表面结构，但表面结构的结构，是潜藏于标示图之下的未标示图。因此在这个'深度结构'中的行为人定位，才真正代表着他/她在群体中的位置。"

A. 结构等价



B. 自形等价

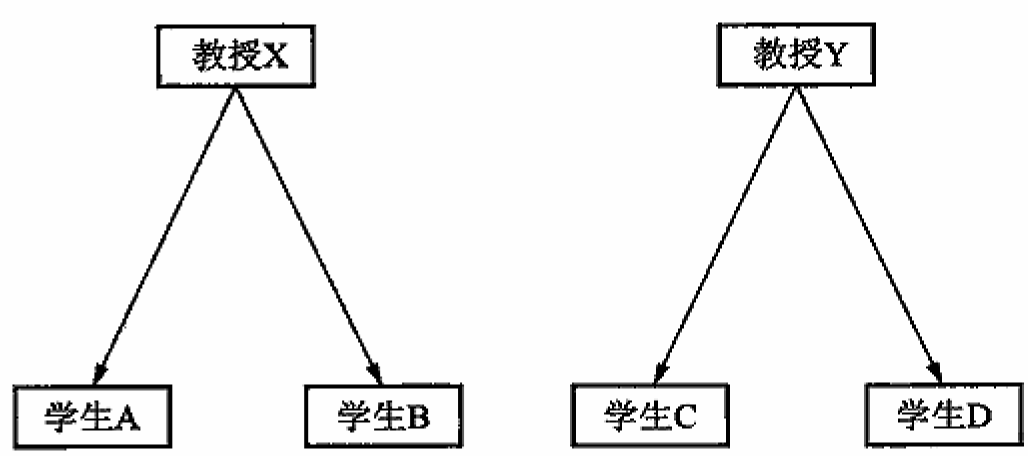


图 5.1 结构等价与自形等价

通过放宽结构等价标准，自形等价法大大便利了对许多社会理论进行实证研究。Borgatti 和 Everett(1992)认为，有几项采用结构等价操作理论的研究，如果采用自形等价进行分析会更好。例如，他们用此方法回应了 Burt(1979)的研究计划。这个研究计划将一个经济体的行业和部门定义为生产相似类型产品、在跨组织网络内占据单一位置的单位集合。Borgatti 和 Everett(1992:21)则宣称，结构等价性单位，如果按照定义，必须要从同一供应商购买原料，产品

必须要销售给同一代理商，实际上这不太可能组成任何符合该定义的部门。但是自形等价单位，则只需要从相似的供应商购买原料，将产品销售给相似的顾客，就能够很好地构成满足要求的行业和部门了。

正规等价是三个常用等价形式中条件最为宽松的一个。它既不要求结构等价中与相同行为人的联结，也不要求自形等价中不可区分的位置。如果行为人与其他也是正规等价的行为人具有同种类型的关系，那么这些行为人就是正规等价的。将这一想法进行概念化的另一条途径是，如果占据一个位置的第一个行为人与占据第二个位置的行为人进行联结，那么如果第二个行为人构成正规等价，则它必须与第二个位置上的其他行为人有着一个相同的联结(White & Reitz, 1983:214)。

所有拥有孩子的母亲都是正规等价，而不论她们孩子数量的多少；反过来也一样。在医院里，医生与他们的患者和护士的关系，也是正规等价的，即使每个医生所属的病人和护士数量不等。正规等价的这种特性使它可能成为社会学家分析社会角色和位置时最为重要的工具。

自形等价和正规等价这两种方法都要求某对行为人与其他在同一种关系具有结构等价性的行为人联系。但是，自形等价和正规等价之间的区别有时也是模糊不清的。自形等价要求未标示图之间必须具有严格的可替代性，但是正规等价并没有要求子图之间具有完全的可替代性。

为了解这种不同，图 5.2 描述了一个分等级的组织机构图。这个图由 4 个垂直的层次构成，其中的联结是监督关系。CEO 监督 3 个执行经理(A、B、C)，而这 3 个执行经理又监督下面 4 个中层管理者(D、E、F、G)，再下层就是一些基层的雇员(从 H 至 N)。如果我们忽略雇员层，那么执行经理人 B 和 C 就是结构等价，因为他们两个都监督着同样的中层管理者(F 和 G)。但是，A 与 B、C 不是结构等价的，因为 A 所监督的管理者不同。然而，这三个执行经理是正规等价型的，因为每一个人都监督着相同数量的中层管理者(2 名)。如果我们考虑所有的等级层次，那么 B 和 C 也是自形等价型的，因为如果标示被移除，他们的子图都是可以互相替代的。但是，A 的子图不能被 B 和 C 的子图所替代，因为 A 的 2 个中层管理者只监督 3 个基层雇员，而 B 和 C 的子图每个都有 4

个雇员。

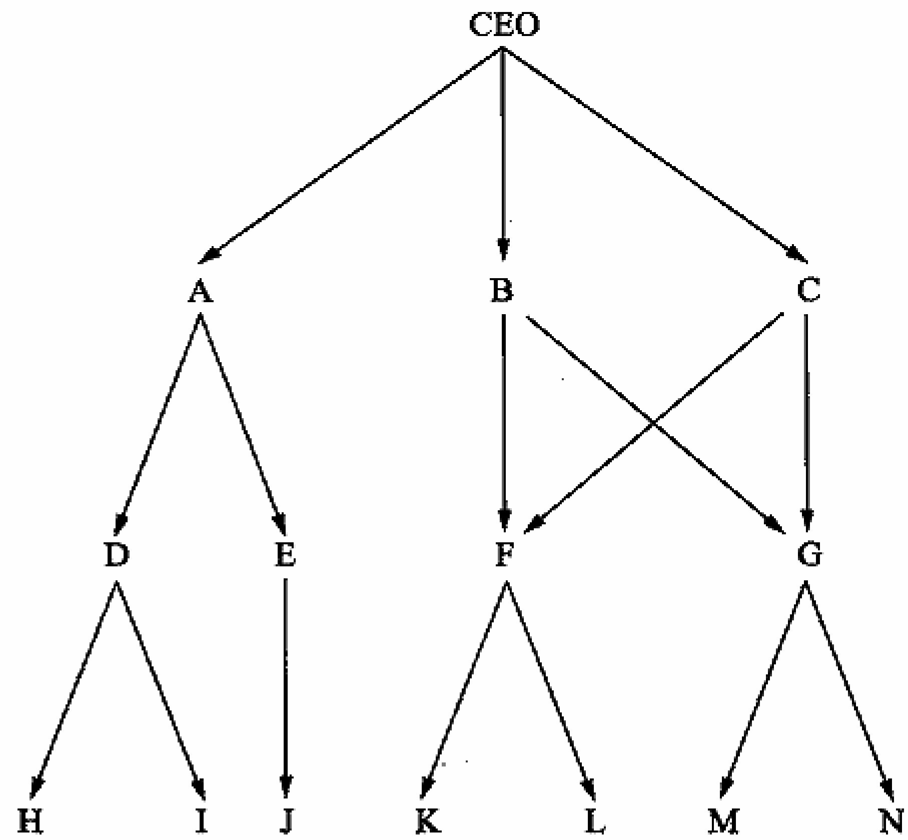


图 5.2 组织等级图的结构等价、自形等价和正规等价

如果仅仅只看两个最低层的关系，那么 4 个中层管理者没有一个是结构等价的，因为他们都监督着不同的基层雇员。然而，其中 3 个中层管理者（D、F 和 G）是自形等价的，因为他们各自下属 2 个雇员，一旦移除标签，是可以完全互相替代的。而 E 不同，它只管理 1 个雇员。但是，如果按照监督至少 1 个雇员的标准，所有的中层管理者都符合正规等价标准。图 5.2 清楚地显示，结构等价是最严格的形式，而正规等价是最宽松的形式，自形等价则处于中间。对于确认一般化的网络社会角色，正规等价似乎是一个非常灵活的方法，它宽泛地定义一个社会系统中与其他位置具有相似结构性关系的行为人阶层或类别。

Logit 模型（p* ）

大多数网络分析方法都是描述性的，试图通过数据简化技术去展现一些潜在的社会关系结构，或者通过代数运算来归纳出网络关系的特征。而 p* 模型，

这个重要的统计方法，它从对问题的描述提升到探索因果关系，即试图将对偶关系的存在作为个体和图论层次的解释因子的函数。Wasserman 和 Pattison (1996)最早提出 p* 模型。这个模型源于 Markov 随机图的早期版本(Frank & Strauss, 1986)——有向图的 log-linear 模型(p1)(Holland & Leinhardt, 1981)，而且还运用了伪似然估计的运算法则(Strauss & Ikeda, 1990)。在随后的发展中，p* 模型延伸至分析多变量关系(Pattison & Wasserman, 1999)和赋值关系(Robins, Pattison & Wasserman, 1999)。p* 模型的上述两个运用描述了数据结构及其结果的解释(Anderson, Wasserman & Crouch, 1999; Crouch & Wasserman, 1998)，并且提供了一个容易获取的介绍(Wasserman & Robins, 2005)。这一部分将简要地讨论基本 p* 模型，并示范运用于一个拟制的网络(Crouch & Wasserman, 1998:87—101)。虽然这个方法可以运用于多变量关系和赋值图的分析，但这里为简化的缘故，我们只处理一个单一关系的二进制有向图。有兴趣的读者如果需要进一步学习，可以查阅上述有关文献。

p* 模型与 logistic 回归密切相关。logistic 回归分析二分因变量(作为二进制变量进行编码：$Y=1$ 或 $Y=0$)，并假定这些变量是二项式分布。如果最小二乘回归法用于建立作为自变量的线性函数的 Y 的概率模型，那么会产生两个主要的问题：(1)一些预测概率可能会高于 1.0 或低于 0.0，而这些明显是没有意义的概率值；(2)违反了关于误差项的核心假设，特别是违反了等方差(同方差性)且正态分布的假设，以及误差项与自变量无关的假设。为了避免这些问题，logistic 回归模型将概率转换为 logits，具体如下：

$$\text{logit}(Y)=\ln\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \qquad [5.1]$$

利用适当的转换，β 参数可以被解释为自变量对于 Y 的 logit、odds 或概率的效应。odds 转换是最普遍的，因为它比 logit 更直接，比概率方法更简单(Pampel, 2000)。为了获得自变量对 odds 的效应，在上述方程两头以 e 为底取幂：

$$\frac{\Pr(Y=1)}{\Pr(Y=0)}=\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)=e^{\beta_0}e^{\beta_1y_1}\cdots e^{\beta_ky_k} \qquad [5.2]$$

这个方程表示，自变量对因变量为 1 相对于因变量为 0 的 odds 的影响是可积的。第 i 个自变量对因变量 odds 的净影响是对回归系数取 e 次幂(e^{β_k})。

p* 模型是应用到社会网络数据的 logistic 回归。对于 g 个行为人之间存在二元有向关系的单一关系网络，$g \times g$ 阶矩阵 X 中的单元(i, j)是个二进制数，

$$x_{i,\,j}=\begin{cases}1 & \text{如果 } i \rightarrow j \\ 0 & \text{其他}\end{cases}$$

Wasserman 和 Pattison(1996)提出 3 个附加矩阵：(1) $X^{+}_{i,\,j}$ 表示从 i 到 j 强行存在的关系链；(2) $X^{-}_{i,\,j}$ 表示从 i 到 j 强行缺失的关系链；(3) $X^{C}_{i,\,j}$ 表示从 i 到 j 的附加关系链。使用这些矩阵，从 i 到 j 的关系链存在且依赖于附加关系链的概率如下所示：

$$\Pr(x_{i,\,j}=1 \mid x^{C}_{ij})=\frac{\Pr(x=x^{+}_{ij})}{\Pr(x=x^{+}_{ij})+\Pr(x=x^{-}_{ij})}$$

$$=\frac{\exp\{\theta Z(x^{+}_{ij})\}}{\exp\{\theta Z(x^{+}_{ij})\}+\exp\{\theta Z(x^{-}_{ij})\}} \qquad [5.3]$$

其中 θ 表示估计的效应参数的向量，$Z(x^{+}_{ij})$ 和 $Z(x^{-}_{ij})$ 分别表示不同自变量或解释变量的原始数据组成的向量。

从 i 到 j 的关系链存在相对于关系链缺失的 odds 是：

$$\frac{\Pr(x_{i,\,j}=1 \mid x^{C}_{ij})}{\Pr(x_{i,\,j}=0 \mid x^{C}_{ij})}=\frac{\exp\{\theta Z(x^{+}_{ij})\}}{\exp\{\theta Z(x^{-}_{ij})\}} \qquad [5.4]$$

两边取自然对数，然后把方程转换成：

$$\log\left[\frac{\Pr(x_{i,\,j}=1 \mid x^{C}_{ij})}{\Pr(x_{i,\,j}=0 \mid x^{C}_{ij})}\right]=\theta[Z(x^{+}_{ij})-Z(x^{-}_{ij})] \qquad [5.5]$$

这个称为 p* 的方程包括估计的效应参数 θ 向量，和当强行把变量 $x_{i,\,j}$ 从 1 变成 0 时原始数据构成的向量。

下面我们说明一下几个关于 p* 模型的问题。首先，分析的单元是一个全幅网络中的对点；因此，对于 N 个行为人组成的有向图，分析单元总数是

N^2-N；第二，虽然模型看起来像是估计对点的特征，但这实际上是整个图的模型。而且，p* 模型估计观察现在的图作为不同网络特征的函数的条件概率，将这些不同的网络特征(如选择、相互关系和传递性)作为自变量。第三，这些自变量的值是当从 i 到 j 的给定关系链由存在变成缺失时，整个图的变量的合计值的差。例如，假设在某个特定图中相互关系链的总数是 5，如果两个个体之间的关系链存在，如从行为人 4 到 5，然后把从 4 到 5 的关系链变成缺失，那么这个图中的一条相互关系链也将消失。因此，对于两个行为人 (4↔5) 的自变量"相互关系"的值为 1。p* 模型实际上估计的是，如果关系链(x_{ij})是随机形成的，且每个有向关系链发生的概率相互独立时，观察到现在的图形的概率。

为了说明 p* 模型，我们使用 Crouch 和 Wasserman(1998)中的二进制有向图。图 5.3 显示了 3 个政府研究机构(圆圈表示)和 3 个私立 R&D 实验室(正方形表示)之间的"项目支持"关系。例如，政府研究机构 1 和 2 之间有相互支持关系。虽然研究机构 3 给私立 R&D 实验室 4 提供项目支持，但这个关系链不是互惠的，实验室 4 对研究机构 3 不提供项目支持。

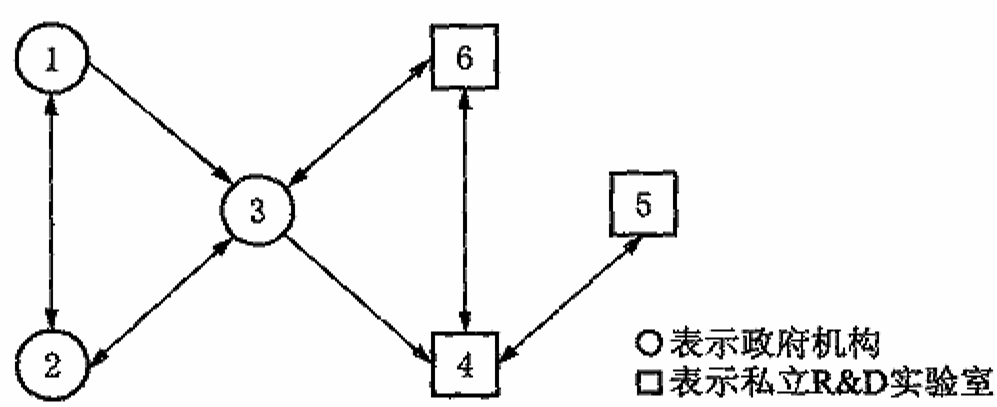


图 5.3 3 个政府研究机构和 3 个私立 R&D 实验室之间"项目支持"的有向图

项目支持似乎更常发生在同一类机构中，也就是说，更常发生在各政府研究机构之间，或各私立实验室之间，而不是机构和私立实验室之间。为了模型化这种"同类"假设和有关关系链存在的其他网络关系的假设，Crouch 和 Wasserman(1998)建模用了 5 个 p* 参数：(1)整体选择度(θ)；(2)同一类型内的不同选择(θ_w)；(3)相互关系(ρ)；(4)同一类型内的不同相互关系(ρ_w)；(5)传递性(τ_T)，关系是可传递的是指当每次从 i 到 j 存在直接关系链，从 j 到 k 存在直接

关系链，那么从 i 到 k 也存在关系链 ($i \to j$, $j \to k$, $i \to k$) (Wasserman & Faust, 1994:164)。在实际的例子中，我们感兴趣的是项目支持是否更常发生在同一类组织之间，互惠的项目支持在同一类组织是否比在不同类组织之间更常发生。因此，要估计的参数向量为：

$$\theta = \{\theta, \theta_w, \rho, \rho_w, \tau_T\}$$

估计有向图中所有两个行为人之间解释变量的向量，包括计算当 i 和 j 之间的关系链从 1 变成 0 时，原始数据的向量 $Z(x)$ 的变化。其中

$Z_\theta = \sum X_{ij}$ 表示解释变量"选择"，

$Z_{\theta_w} = \sum X_{ij}\delta_{ij}$ 表示变量"同一类型内部选择"，

$Z_\rho = \sum_{i<j} x_{ij}x_{ji}$ 表示变量"相互关系"，

$Z_{\rho_w} = \sum_{i<j} x_{ij}x_{ji}\delta_{ij}$ 表示变量"同一类型内部相互关系"，

$Z_{\tau_T} = \sum_{ijk} X_{ij}X_{jk}X_{ik}$ 表示变量"传递性"。

指标变量 δ_{ij} 是个二进制指标，当 i 和 j 是相同类型时，δ_{ij} 为 1，否则为 0。对于有 g 个节点的有向图，p* 模型中有序数对的总数是 $g(g-1)=g^2-g$。因此，6 个组织构成 30 个有序数对。由于篇幅所限，这里没有呈现所有的 p* 模型参数。有兴趣的读者可以参考 Anderson(1999)、Wasserman 和 Pattison(1996)的相关文献。

表 5.1 列出了用于估计 30 个有序数对的 p* 参数的输入数据。下面的例子说明了是如何把这 5 个公式应用到一个有序对偶关系 ($2 \to 3$) 中来计算输入数据的值的，这个有序对偶关系是表 5.1 中第 7 行从行为人 2 到行为人 3 的有向关系链。首先，通过当从行为人 i 到 j 的关系链存在时的关系链总数减去当从行为人 i 到 j 的关系链缺失时的关系链总数，从而得到变量"选择"：($\sum X_{ij}^{+} - \sum X_{ij}^{-}$)。在这个例子中，对于每对对点，变量"选择"的值为 1。因此，变量"选择"实际上是个常数，类似于模型中的截距。第二，同类组织内部的选择(从行为人 2 到行为人 3)为 1，因为如果从行为人 2 到行为人 3 的关系链存在，则同类

表 5.1 p* 数据

观察个体	*i*	*j*	关系链	选择	同类间选择	相互关系	同类间相互关系	传递性
1	1	2	1	1	1	1	1	2
2	1	3	1	1	1	0	0	3
3	1	4	0	1	0	0	0	1
4	1	5	0	1	0	0	0	0
5	1	6	0	1	0	0	0	2
6	2	1	1	1	1	1	1	1
7	2	3	1	1	1	1	1	2
8	2	4	0	1	0	0	0	2
9	2	5	0	1	0	0	0	0
10	2	6	0	1	0	0	0	3
11	3	1	0	1	1	1	1	3
12	3	2	1	1	1	1	1	1
13	3	4	1	1	0	0	0	3
14	3	5	0	1	0	0	0	2
15	3	6	1	1	0	1	0	2
16	4	1	0	1	0	0	0	0
17	4	2	0	1	0	0	0	1
18	4	3	0	1	0	1	0	3
19	4	5	1	1	1	1	1	0
20	4	6	1	1	1	1	1	1
21	5	1	0	1	0	0	0	0
22	5	2	0	1	0	0	0	0
23	5	3	0	1	0	0	0	1
24	5	4	1	1	1	1	1	0
25	5	6	0	1	1	0	0	3
26	6	1	0	1	0	0	0	1
27	6	2	0	1	0	0	0	3
28	6	3	1	1	0	1	0	1
29	6	4	1	1	1	1	1	2
30	6	5	0	1	1	0	0	3

组织内的选择总数为 9；当这个关系链不存在时，同类组织内的选择总数为 8。行为人 3 和 6 之间的相互关系链和从行为人 3 到行为人 4 的有向关系链没有计

算在内是因为这两对对点是不同类型组成的。第三，这个有向图存在 5 个相互关系链 (1↔2, 2↔3, 3↔6, 4↔6, 4↔5)。当从行为人 2 到行为人 3 的有向关系链从存在变成缺失时，相互关系链的总数减少到 4。因此，从行为人 2 到行为人 3 的相互关系变量值为 $5-4=1$。第四，这个有向图存在 4 条同类组织间的相互关系链 (1↔2, 2↔3, 4↔6, 4↔5)，不包括行为人 3 和行为人 6 之间的相互关系链，因为这对对点是不同类型的。因此，当从行为人 2 到行为人 3 的有向关系链从存在变成缺失时，同类组织的相互关系链总数从 4 减少到 3，即减少了 $4-3=1$。第五，这个有向图有 5 种传递的关系 [(1 → 2, 2 → 3, 1 → 3), (1 → 3, 3 → 2, 1 → 2), (2 → 1, 1 → 3, 2 → 3), (3 → 6, 6 → 4, 3 → 4), (3 → 4, 4 → 6, 3 → 6)]。当从行为人 2 到行为人 3 的有向关系链从存在变成缺失时，2 种传递的关系 (1 → 2, 2 → 3, 1 → 3) 和(2 → 1, 1 → 3, 2 → 3) 将消失。因此，这个传递性变量的值为 2。总的来说，对于有序对点 (2 → 3) 的 5 个计算如下所示：

选择的变化：$\sum X_{ij}^{+} - \sum X_{ij}^{-} = 1$。

同类组织间选择的变化：$\sum X_{ij}^{+}\delta_{ij} - \sum X_{\circ ij}^{-}\delta_{ij} = 9-8=1$。

相关关系链的变化：$\sum_{i<j}^{+} x_{ij}x_{ji} - \sum_{i<j}^{-} x_{ij}x_{ji} = 5-4=1$。

同类组织之间相互关系的变化：$\sum_{i<j}^{+} x_{ij}x_{ji}\delta_{ij} - \sum_{i<j}^{-} x_{ij}x_{ji}\delta_{ij} = 4-3=1$。

传递性的变化：$\sum_{ijk}^{+} X_{ij}X_{jk}X_{ik} - \sum_{ijk}^{-} X_{ij}X_{jk}X_{ik} = 5-3=2$。

表 5.1 中的数据可以用 SPSS, SAS 和 STATA 等统计软件中的 logistic 回归来分析。因变量是二分变量“关系链”，对自变量——选择，同类组织间的选择，相互关系，同类组织间的相互关系和传递性——进行回归。回归结果如表 5.2 所示。在这个模型中，只有一个解释变量——相互关系链——是显著的。包括从 i 到 j 的有向关系链在内的相互关系链每增加一条，i 发送一条关系链到 j 的 odds 增加将近 41 倍 ($\exp(3.737)-1=40.97$)。控制其他解释变量后，一条相互关系链存在的 odds 是不存在的 odds 的 41 倍多(Anderson et. al.,

表 5.2 网络关系链的非标准化 logistic 回归

变　量	模型 1	变　量	模型 1
选择	−2.208 (1.197)	传递性	−0.408 (1.197)
同类间选择	2.741 (2.183)	卡方(自由度)	19.488*** (4)
相互关系	3.737* (1.829)	−2 log likelihood	20.893
同类间相互关系	−1.588 (2.921)		

注:括号里的是标准误。* 表示 $p<0.05$,** 表示 $p<0.01$,*** 表示 $p<0.001$(双尾检验)。

1999:54)。因此,这个有向图中关系链是互惠型的可能性远远大于不是非互惠型的可能性。然而,互惠和选择对于不同类型的组织是不存在差异的,同类组织内的选择和同类组织内的相互关系都是不显著的。这个结果不支持项目支持在同类组织内比不同组织间更可能发生的假设。

在社会网络分析方法中,p* 模型的发展标志着一个非常重要的进步。从此,社会网络分析不仅仅只是描述行为人之间的关系链,还可以解释行为人之间的关系链。模型也能清晰地模拟对图层特征和个体变量的关系链的影响。p* 模型及其类似的模型在网络分析研究人员用来分析实际网络问题时将大有用处。许多实际介绍 p* 模型的教科书将有助于普及这种方法在网络研究人员中的应用。

隶属网络

隶属网络,也称为成员资格网络,表示一组行为人对一系列社会事件的参与(Wasserman & Faust, 1994:291—343)。人们可以通过对社会活动的联合参与互相联结,或者因他们共同加入某一组织而互相联系。反过来说,社会事件因其共有的参与人而互相联结。隶属网络分析形象地展示了行为人和事件

是如何同时相互联系的。

一个隶属网络由两种类型的节点(一组行为人和一系列事件)和每个节点类型之间的一系列关系构成。因此,隶属网络是一种二模网络,这与连接行为人与行为人的一模网络不同。关于隶属网络的实际研究非常多(Wasserman & Faust, 1994:196)。在第3章,我们讨论了Freeman等人(1987)关于大学教职工和学生参加9个座谈会的研究。研究者使用隶属网络来表示这些关系,其中参与人是行为人,座谈会是事件。其他一些大家熟悉的例子,有属于志愿者组织的成员、参加抗议活动的社会运动活动者、创立战略联盟的企业,以及签订军事协议的国家等。对二模网络的研究常常具有双重目的:揭示行为人之间通过共同参与事件而形成的关系结构;揭示事件吸引共同的参与人的关系结构。

隶属矩阵和二分图 一个二进制隶属网络可以由一个隶属矩阵来表示,该矩阵记录 g 个行为人在 h 个事件中的参与和缺席。因此,它的维度分别是 g 行和 h 列。如果行为人 i 参与事件 j,那么矩阵中第 i 行第 j 列的单元格的元素等于1,否则元素为0。将这个二进制隶属矩阵命名为A,它的 x_{ij} 值符合下述条件:

$$x_{i,j}=\begin{cases}1 & \text{如果行动者 } i \text{ 与事件 } j \text{ 有关}\\ 0 & \text{其他}\end{cases}$$

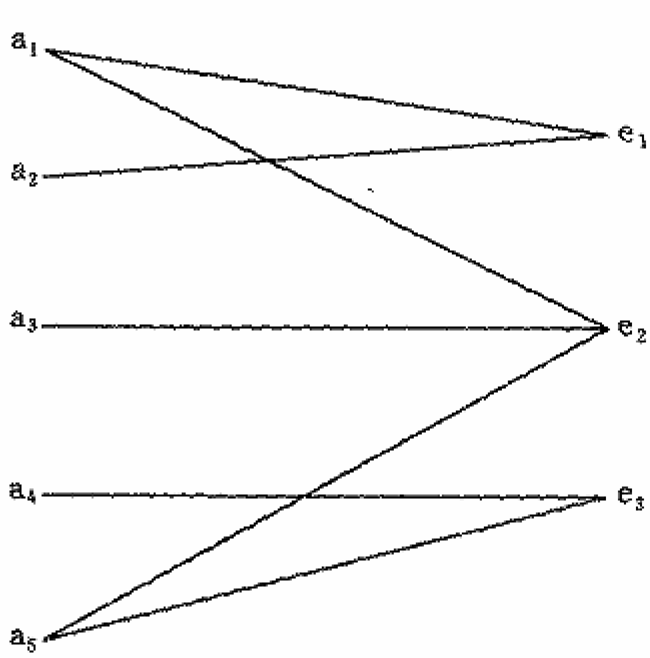


图5.4 隶属网络的二分图

矩阵A的行总数,也称为行的边际值,记录每一个行为人所参与事件的总数 $\sum_{j=1}^{h}x_{i,j}$。列的边际值($\sum_{j=1}^{g}x_{i,j}$)表示参与每一个事件的行为人的数量。

隶属网络也可以表示为二分图,其中无向线连接排列在图表一端的行为人和排列在图表另一端的事件。重要的是,二分图不允许行为人之间的连

线和事件之间的连线。二分图矩阵(B)在其行和列中包含行为人和事件。假设一个隶属网络有 g 个行为人和 h 个事件，那么 B 的维度为 $(g+h)\times(g+h)$。

图 5.4 是一个隶属网络的二分图，其中有 5 个行为人和 3 个事件。线条表示哪一位行为人参与了哪一个事件，同时也表示哪一个事件吸纳了哪一位行为人。虽然在这个图中行为人之间或事件之间没有直接的连线，但是它表明了行为人是如何通过他们共同隶属的具体事件而相互连接的。例如，行为人 1 和行为人 2 都参与了事件 1，而行为人 1、3、5 则因事件 2 互相联系在一起。

表 5.3 是对应于图 5.4 的一个二分矩阵，单元格的值为 1 表示一个行为人和一个事件相连。

表 5.3　图 5.4 对应的二分矩阵

	a1	a2	a3	a4	a5	e1	e2	e3	小计
a1	—	0	0	0	0	1	1	0	2
a2	0	—	0	0	0	1	0	0	1
a3	0	0	—	0	0	0	1	0	1
a4	0	0	0	—	0	0	0	1	1
a5	0	0	0	0	—	0	1	1	2
e1	1	1	0	0	0	—	0	0	2
e2	1	0	1	0	1	0	—	0	3
e3	0	0	0	1	1	0	0	—	2
小计	2	1	1	1	2	2	3	2	—

行为人 1 有两个 1 值，表明它参与了 2 个事件(事件 1 和事件 2)。在右上角象限区内的 5×3 子矩阵，由行的 5 个行为人和列的 3 个事件构成，是一个隶属矩阵 A。左下角象限区是 A 的转置，表示为 A′，其维度是 3×5，且 $x_{i,j}=x_{j,i}$。转置的隶属矩阵表明，行中的哪一个事件吸引了列中的哪一位行为人。二分矩阵的另外 2 个象限的单元格的值总是为 0，因为在二模数据中，行为人之间不直接连接，事件之间也是如此。二分矩阵可以简化为如下形式：

$$X^{A,E}=\begin{bmatrix}0 & A\\ A' & 0\end{bmatrix} \qquad [5.6]$$

在表格的边际上(并非矩阵的一部分),行的总数等于列的总数。行为人总数表明它参与事件的数量,而事件总数则表明参与行为人的数量。

将两个处于不同顺序的子矩阵(A 和 A′)相乘,便得到额外的信息,即行为人之间的关系和事件之间的关系,而这在二分矩阵中是没有的。首先,X^A 是一个行为人对点共同出席的赋值的对称矩阵,由两个矩阵相乘获得:

$$X^A = AA' \qquad [5.7]$$

一般来说,隶属网络 A 是一个 $g \times h$ 矩阵,而它的转置 A′是一个 $h \times g$ 矩阵。因此,X^A 总是一个 $g \times g$ 矩阵,它的非对角线单元格的值是行为人 i 和 j 共同参与的事件数量。A 的对角线元素表示每一个行为人参与的事件数量。其次,X^E 是一个事件对点共同参与的赋值的对称矩阵,由两个矩阵相反相乘获得:

$$X^E = A'A \qquad [5.8]$$

这个 $h \times h$ 矩阵的非对角线单元格的值,是同时参与事件 i 和 j 的行为人数量。而 A′矩阵的对角线元素,是参与每一个事件的行为人数量。

案例数据的共同出席矩阵图如下所示:

$$X^A = AA' = \begin{pmatrix} 110 \\ 100 \\ 010 \\ 001 \\ 011 \end{pmatrix} \times \begin{pmatrix} 11000 \\ 10101 \\ 00011 \end{pmatrix} = \begin{pmatrix} 21101 \\ 11000 \\ 10101 \\ 00011 \\ 10112 \end{pmatrix} \qquad [5.9]$$

5 个行为人和 3 个事件形成的 X^A 是一个 5×5 的对称矩阵,其非对角线单元格的值表示每一对点参与事件的数量。如第一行所示,行为人 1 与行为人 2、3 和 5 各参与 1 个事件,但未与行为人 4 参与事件。对角线表明,行为人 1 和 5 各参与 2 个事件,而其他人则只参与了 1 个事件。

案例数据的共同参与矩阵如下所示:

$$X^E = A'A = \begin{pmatrix} 11000 \\ 10101 \\ 00011 \end{pmatrix} \times \begin{pmatrix} 110 \\ 100 \\ 010 \\ 001 \\ 011 \end{pmatrix} = \begin{pmatrix} 210 \\ 131 \\ 012 \end{pmatrix} \quad [5.10]$$

这个 3×3 的对称矩阵的非对角线单元格表示参与每对事件的行为人数量。事件 1 和事件 2 只有一个共同的行为人(其中行为人 1 同时参加了这 2 个事件),但是事件 1 和事件 3 没有共同的行为人。对角线值反映,事件 2 吸引了 3 个行为人(行为人 1、3 和 5)。

在分析的网络层次上,隶属的平均率值是可以从两个乘积矩阵值计算出来的。因为 X^A 的对角线元素是每一个行为人参与事件的数量,因此将它们加起来,然后除以行为人(g)的数量,即得到平均的行为人参与率 $\overline{X}^A = \frac{\sum_{i=1}^{g} x_{i,i}}{g}$。在这个例子中,平均有 $\left(\frac{7}{5}\right) = 1.40$ 个行为人参与每一个事件。同样,将 X^E 的对角线值相加,除以事件(h)总数,即得到平均的事件参与率 $\overline{X}^E = \frac{\sum_{i=1}^{h} x_{i,i}}{h}$。在这个例子中,平均的事件参与率是 $\left(\frac{7}{3}\right) = 2.33$ 个行为人。

隶属网络的密度和中心度　密度和中心度是非常重要的基本网络特征,它们也能够适用于隶属网络。正如在第 4 章所讲的,密度测量反映二进制图中联结出现的比例,或者反映赋值图中可观察到的线的平均值。同样地,对隶属网络密度的阐释也依赖于该网络是否为二进制图或赋值图(Wasserman & Faust, 1994:316)。

对于共同参与的 $g \times g$ 对称矩阵 X^A 来说,其非对角线值是每对行为人参与事件的数量,其密度计算方程如下:

$$D^A = \frac{\sum_{i=1}^{g}\sum_{j=1}^{g} X_{ij}^A}{g(g-1)/2} (i < j) \quad [5.11]$$

在方程 5.11 中，分子表示共同出席矩阵中上三角阵中所有值的总和（即对角线以上的部分，因为下部分三角的值是相同的）。因为我们不可能认为一个行为人与他自己共同参与某一事件，所以我们排除了对角线上的值。分母是无序对点的所有数量，当然也排除了对角线上的值。对于共同参加的 $h \times h$ 对称矩阵 X^E 来说，其非对角线上的值是每一个事件中参与的行为人数量，其密度计算如下所示：

$$D^E = \frac{\sum_{i=1}^{h} \sum_{j}^{h} X_{i,j}^{E}}{h(h-1)/2}(i < j) \qquad [5.12]$$

在这个例子中，$D^A = 5/10 = 0.50$，意味着每对行为人共同参与了平均半个事件。从事件角度来看，$D^E = 2/3 = 0.67$，表明平均每对事件吸引了三分之二个行为人。由于乘积矩阵单元格的值可能大于 1，因此密度值也可能大于 1.0。因此，每一个密度值不应解释为一个比例，而应该分别解释为每两个行为人所出席的事件平均数（D^A），或每两个事件中的平均参与人数（D^E）。

社会网络分析在行为人的分析层面研究中心度，在图的分析层面研究中心度，这种研究已经进行了数十年（Freeman，1979；Wasserman & Faust，1994，第 3 章）。我们在第 4 章讲过，行为人中心度测量一个网络之内行为人的重要性或显性。根据研究者定义重要性和显性的不同，可以将中心度分为 4 种主要的类型：程度、亲密性、中间性和特征向量中心度。程度中心度反映行为人在网络中活动的积极程度；亲密性中心度测量一个行为人通过网络内最短路径与其他行为人联结的程度；中间性中心度捕捉一个行为人对网络其他行为人之间的信息或资源流的居间程度；特征向量中心度反映行为人与网络中的其他中心行为人的联系程度。Faust（1997）讨论了这 4 种中心度测量方法在隶属网络中的运用，但在此由于文本篇幅的限制，我们只讨论隶属网络中程度中心度的使用。

一般认为，程度中心度包含有向联结的总数，因此隶属网络的行为人程度中心度便是行为人联系的总数，即第 i 个行为人出席所有事件时所接触到的联结数量，由共同出席的矩阵 X^A 的第 i 个行值相加所得：

$$C_D^A(a_i) = \sum_{j=1}^{g} x_{ij}^A (i \neq j) \qquad [5.13]$$

在这个例子中，行为人 1 分别与行为人 2、3 和 5 一起参与了 3 个不同事件，因此它的程度中心度是 3。行为人 2 的程度中心度更低，因为它只与行为人 1 参与过第 1 个事件。

同样地，隶属网络中一个事件的程度中心度是事件联系的总数，即第 j 个事件通过所有行为人的参与而得到的联结数量，由共同参与矩阵 X^E 的第 j 个列值相加所得：

$$C_D^E(e_i) = \sum_{j=1}^{h} x_{ij}^E (i \neq j) \qquad [5.14]$$

在这个例子中，3 个事件的程度中心度分别为 1、2、1。事件 2 的程度中心度值最高，因为通过至少一个共同的行为人的参与，从而与事件 1 和事件 3 都有了联结。而事件 1 和事件 3 则只与事件 2 共享行为人。

点阵

上面所讨论的隶属网络，一个明显优势就在于可以看见二模数据网络结构的三种类型：行为人—事件关系及其转置，通过共同参与事件而构成的行为人之间的关系，通过吸引共同行为人而构成的事件之间的关系。然而，隶属网络的二分图形式不能揭示行为人和事件节点二者是如何相关的。Galois 点阵提供了一个二模数据的一般表述，使得同时查看两个模式之间的关系成为可能，为观察网络的结构特征提供了有价值的研究视角（Freeman & White, 1993; Wasserman & Faust, 1994:326—342）。在实证研究中，该方法运用于下列关系的研究：中国革命年代的村民和政治冲突（Schweizer, 1996），济贫机构为贫民提供的救济和待遇（Mohr & Duquenne, 1997），以及 1992 年巴西弹劾运动中的组织和事件（Mische & Pattison, 2000）。

Galois 点阵　Galois 点阵，因 19 世纪法国数学家 Évariste Galois 的论著而得名，也称为 Galois 连接，是两个有序集合之间的连接。在网络运用中，Galois

点阵有两个非空集合：行为人集合 A 和事件集合 E。这两个集合通过二模隶属网络连接，表示行为人是如何与事件相连的（反之亦然）。因此，一个 Galois 点阵由一个三元组（A，E，I）构成，其中 I 是矩阵 A × E 中的二进制关系。表5.3 中二分矩阵的子矩阵 A 是一个行为人一事件矩阵的实例。现在考虑 P(A) = {A_1，A_2，…}，这是一系列矩阵 A 的集合；P(E) = {E_1，E_2，…}，这是一系列矩阵 E 的集合。I 关系定义为从 P(A)至 P(E)的映射：B→B↑：

$$B\uparrow = \{e \in E \mid (a, e) \in I, \text{所有 } a \in A\} \quad [5.15]$$

这个数学表达式表明，映射图能够识别某一特定行为人或某些行为人所隶属的所有事件。尤其是，↑映射从一系列行为人到一系列事件，其中这些行为人隶属于这些事件（Wasserman & Faust，1994：330）。例如，表 5.3 中的子矩阵 A 表示，行为人 1 隶属于事件 1 和事件 2，而行为人 1 和 2 都隶属于事件 1。

反过来，这个映射能够从 P(E)至 P(A)：$F \rightarrow F\downarrow$

$$F\downarrow = \{a \in A \mid (a, e) \in I, \text{所有 } e \in E\} \quad [5.16]$$

方程(5.16)意味着映射应当确认一个具体事件所吸引的所有行为人。表 5.3 中的子矩阵 A 显示，事件 2 吸引了行为人 1、3 和 5，但事件 2 和事件 3 只吸引了行为人 5。

将两个映射图合并起来，Galois 点阵展现了一系列行为人是如何隶属于一系列事件的。在图表中的每一个点，既表示一系列行为人，也表示一系列事件。为了解释 Galois 点阵，需要在一个一致的方向上移动：(1)为了跟踪出席事件的行为人，只需要向上移动，将点阵从底端升至顶端；(2)为了跟踪吸引了行为人的事件，只需要向下移动，将点阵下降。依照惯例，点阵上面部分涵盖了行为人集合 A 的全部要素，下面部分涵盖了行为人集合 E 的全部要素。图 5.5 展示了表 5.3 当中隶属矩阵的 Galois 点阵图。这个图将 3 个事件标示为 A、B、C，将 5 个行为人标示为 1、2、3、4、5。该点阵底部聚集了最多的事件。沿着点阵向上移动，会逐渐遇到行为人集合的增大和事件集合的缩小。在底端的起始点，包含所有的事件(ABC)，但行为人集是空的(ϕ)，因为没有人参与所有的事件。向上一点，点阵图显示，事件 A 和事件 B 共享行为人 1，事件 B 和事件 C 共享行为

人5。再往上,事件集合变得更小,但行为人集合变大。事件A吸引了行为人1和2,而事件B吸引了行为人1、3和5,事件C吸引了行为人4和5。点阵最顶端罗列了所有的行为人但没有任何事件,这表明没有事件吸引所有的网络行为人。一般来说,如果接收到从事件处下降之线条上行为人是入射的,那么该行为人隶属于该事件。在以事件A和事件B为出发点的线条上,行为人1是入射的,表明行为人1隶属于这两个事件。相反,在从行为人处升起之线条上是入射的事件,吸引了这些行为人。例如,事件C在由行为人5处升起之线条上是入射的,表明事件C吸引了行为人5。事件B收到来自行为人1和5的线条,表明这两个行为人都参与了事件B。

Galois点阵也表示一些事件之间的隶属模型和行为人之间的隶属模型。例如,图5.5显示,行为人2没有参与任何无行为人1参加的其他事件(行为人2只参加了事件A,行为人1也隶属其中)。同样,行为人4只参加了事件C,该事件也吸引了行为人5。相反,行为人1或5都没有限制自己参加同样吸引了行为人2或4的事件。也就是说,对于行为人1或5来讲,参与某个事件并不以其他行为人的参与为条件。从事件的角度来看,图5.5也显示,3个事件集合中的行为人有重叠但不相同,在此意义上3个事件都包含不同的行为人集合。含有更多行为人和事件的更复杂网络,也许会发现一些包容性结构,其中在某一个事件中出现的行为人也出现在其他的事件里(Freeman & White, 1993)。因此,所有3种连接类型——行为人—事件关系、行为人—行为人关系、事件—事件关系——的数据都能够用Galois点阵进行精确重组。

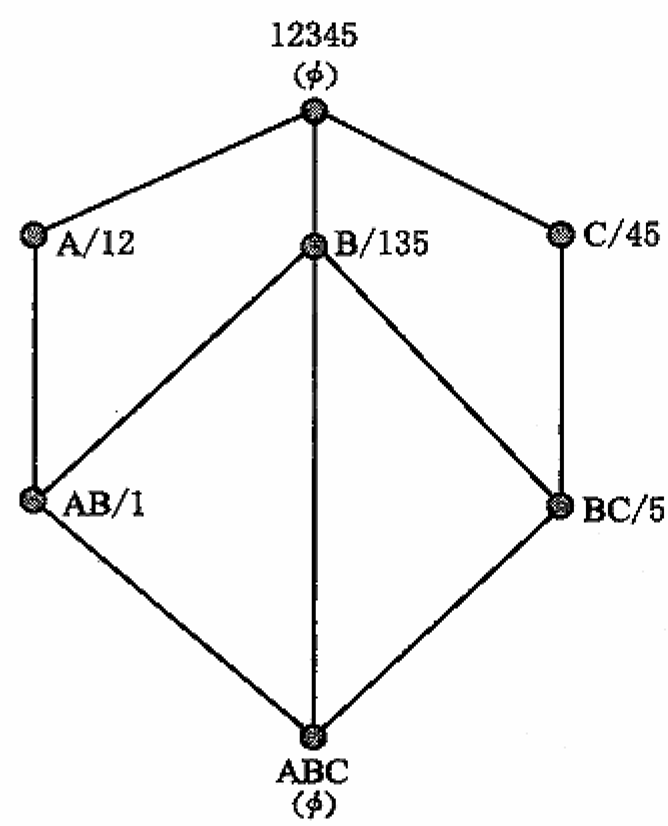


图5.5 隶属网络的Galois点阵

使用Galois点阵对3种关系类型的结构特点进行同步分析,尽管这是一个非常明显的优势,但是它的应用只限于表述小型的二模网络数据集合。关于这

方面,Galois 点阵类似于图,其主要的运用是展示数据,而不是还原数据。大型数据集合,一般包含高度复杂的结构,会迅速搅乱 Galois 点阵的可视性。即使利用简化符号,如 Freeman 和 White(1993)所建议的那样,观察者依然很难看清楚复杂的 Galois 点阵的图像。

网络小团体的 Galois 点阵 网络分析者认为,一些统计的或代数的数据还原方法可以被用于简化 Galois 点阵的可视化图像(Duquenne, 1996; White, 1996)。Freeman(1996)认为,Galois 点阵能够帮助展示社会行为人之间的小团体。经典的 Luce-Perry 小团体定义假定存在一个关于社会关系 R 的二进制对称平方矩阵(A^2)。小团体 C 是一个含有三个或多个行为人的最大子集,其中每对行为人都由 R 连接。"最大"意味着没有小团体能涵盖在一个更大的小团体中。然而,在实践中,小团体往往太小、太多或者太过于重叠,以至于不能反映社会群体的直观结构(Freeman, 1996:174)。Galois 点阵方法运用于小团体结构的可视化处理,需要用小团体集替代事件集。之前我们讲到过,Galois 点阵由一个三元组(A, E, I)构成,其中 A 是行为人集,E 是事件集,I 是 A×E 中的二进制关系。如果将 Galois 点阵应用于小团体,那么 Galois 点阵将被重新定义为另一个三元组(A、C 和 I),其中 A 和 I 的意义不变,C 表示小团体的集合。

遵循事件的 Galois 点阵的相似布局,小团体 Galois 点阵将较大的小团体集合放置在下部,将较大的行为人集合放置在上部。假定图 5.5 现在表示的是一个小团体 Galois 点阵,而不是一个事件 Galois 点阵,那么在底部有 3 个小团体 ABC,其零值表示没有行为人属于这 3 个小团体。沿着点阵往上走,我们可以发现,小团体越少,行为人集合越大。行为人 1 同时属于小团体 A 和 B,而行为人 5 同时属于小团体 B 和 C。继续往上走,可以看到小团体 A 包含行为人 1 和 2,小团体 B 包含行为人 1、3 和 5,而小团体 C 包含行为人 4 和 5。行为人的全体集合位于图的顶端,是个空集,表示没有一个小团体同时包含所有 5 个行为人。

Freeman(1996)讨论过小团体网络 Galois 点阵的几个重要结构特征。首先,两个重叠小团体由下降线连接,交会于点阵图低处某个标签点,而两个非重叠小团体则仅在无标签泛下限处与空集(ϕ)连接。假设图 5.5 反映的是一个小团体网络的 Galois 点阵,小团体 A 和 B 交会于一个标有行为人 1 的较低点,表

示两个小团体出现重叠，共享了行为人 1。相反，小团体 A 和 C 没有重叠，表示它们的交会点是在无标签的泛下限点。

其次，Freeman(1996)利用多种重要维度(比如链、长度、高度和深度)，对小团体点阵的个人行为人的位置按特点进行了归纳。链是指由一个元素引至另一个元素的全部上升线或下降线所组成的序列。链的长度是链所包含的线条的数量。行为人的高度是从泛下限点上升至该行为人处的链的长度。相反，行为人的深度是从泛上限点下降至该行为人处的链的长度。因此，处于点阵低处的行为人具有较大的深度和较小的高度，而处于点阵顶部的行为人具有较大的高度但是较小的深度。深度较大但高度较小的行为人深嵌在网络中。他们参与多个小团体，但他们与这些小团体的隶属关系独立于其他行为人的隶属关系。从某种意义上来说，他们构成了这个小团体的核心成员。相反，高度较大但深度较小的行为人则往往浮在网络表层。他们与某个小团体的隶属关系往往取决于其他行为人的隶属情况。因此，他们属于这个小团体的边缘人物。将这些观点套用于图 5.5，我们就能够确定行为人 1 和 5 分别是小团体 AB 和 BC 的核心成员，而行为人 2、3 和 4 则是这些小团体的边缘人物。

对应分析 上面讲的点阵分析是运用代数方法展示隶属网络的代表，而对应分析则利用度量技术来联合展现隶属网络中所有的行为人和事件(Faust, 2005; Wasserman & Faust, 1994:291—343)。对应分析主要通过奇异值分解(SVD)的数学方法进行。我们这里只简要地介绍一下 SVD，集中谈论与隶属网络矩阵有直接关系的几个事项(如欲知详情，请参阅 Strang, 1988)。

SVD 是矩阵 A 的分解，其维度有 $g \times h$ ：

$$A = X \wedge Y^T \qquad [5.17]$$

这个方程含有 $\wedge$，它表示奇异值的对角线矩阵 $\{\lambda_K\}$；X 表示大小为 $g \times h$ 的左奇异向量矩阵；Y 表示大小为 $h \times h$ 的右奇异向量矩阵。因此，左奇异向量指矩阵 $g \times h$，右奇异向量指矩阵 $h \times h$。奇异值的数量也被称为秩，通常表示为 W。SVD 利用秩 W 来度量视图中的行为人和事件，将它们排列在 A 的合适位置中。

对应分析中的SVD包括对标准化后的矩阵A的分解。标准化有两种方法:(1)将原始矩阵A的元素除以行和列边际总数乘积的平方根;(2)计算矩阵A和两个对角线矩阵$R^{-\frac{1}{2}}$和$C^{-\frac{1}{2}}$的乘积(Faust, 2005),其中

$$R^{-\frac{1}{2}} = \mathrm{diag}\left(\frac{1}{\sqrt{a_{i+}}}\right) \qquad [5.18]$$

$$C^{-\frac{1}{2}} = \mathrm{diag}\left(\frac{1}{\sqrt{a_{+j}}}\right) \qquad [5.19]$$

将三个矩阵相乘,得到矩阵A的标准化形式,标准化矩阵A的获得也可以通过如下方法:用矩阵A的元素除以行和边际总数乘积的平方根。对应分析包含乘积矩阵的SVD:

$$R^{-\frac{1}{2}}AC^{-\frac{1}{2}} = X \wedge Y^T \qquad [5.20]$$

对应分析获得三个信息集合:行矩阵的g值集合$U = \{u_{ik}\}$,其中$i = 1, 2, \cdots g$, $k = 1, 2, \cdots w$;列矩阵的h值集合$V = \{v_{jk}\}$,其中$j = 1, 2, \cdots h$, $k = 1, 2, \cdots w$;奇异值$\wedge = \{\lambda_K\}$,其中$k = 1, 2, \cdots w$,表示每一个维度的重要性。为了获得一个所有行和列元素的联合展示,对应分析将行为人的值计算为其所隶属的事件加权平均值的函数,将事件的值计算为它的行为人值的加权平均值的函数:

$$\lambda_k u_{ik} = \sum_{j=1}^{h} \frac{a_{ij}}{a_{i+}} v_{jk} \qquad [5.21]$$

$$\lambda_k v_{jk} = \sum_{j=1}^{g} \frac{a_{ij}}{a_{+j}} u_{ik} \qquad [5.22]$$

在上述两个等式里,a_{ij}是原始矩阵A中第i行和第j列的单元格的值。

为了讲解的便利,我们利用前面的隶属网络数据:5个行为人和3个事件。表5.4表示原始版和标准化版的矩阵A。表5.5列出了标准化矩阵A的SVD,它得到3个数据集合u_{ik}、v_{jk}和λ_K。行为人值是其所隶属的事件值的一个函数。同样,事件值是其所吸引的行为人值的一个函数。具体来讲,在第一维度行为人1的值(−0.661)是由该行为人隶属的两个事件的加权平均值除以奇异

表 5.4　隶属网络的矩阵 A

	$e1$	$e2$	$e3$
	隶属矩阵 A		
a1	1	1	0
a2	1	0	0
a3	0	1	0
a4	0	0	1
a5	0	1	1
	标准化矩阵 A		
a1	0.5	0.408	0
a2	0.707	0	0
a3	0	0.577	0
a4	0	0	0.707
a5	0	0.408	0.5

值(λ_1)得出：$\frac{\frac{1}{2}\times(-1.146)+\frac{1}{2}\times 0}{0.866}=-0.661$。同样，在维度 1 中事件 1 的值(−1.146)是由事件所吸引之行为人(1 和 2)的加权平均值得来的：$\frac{\frac{1}{2}\times(-0.661)+\frac{1}{2}\times(-1.323)}{0.866}=-1.146$。那么，你能计算出其他行为人和事件的值吗？

图 5.6 给出了隶属网络的对应分析。括号内的数是 2 个维度的值，等于表 5.5 中的数。X 轴和 Y 轴分别表示第一和第二维度(λ)。在图中，行为人 1 隶属于事件 1 和事件 2，行为人 5 隶属于事件 2 和事件 3，处于 2 个维度的中心，因此位于图表的中央。事件 2 位于第一维度的中心，主要是因为它吸引了大多数行为人(1、3 和 5)，但是它离第二维度的中心要比事件 1 和 3(每个事件只吸引了两个行为人)远一点点。行为人 3 位于第一维度的中心，因为它隶属于中心事件(事件 2)，但是行为人 3 在第二维度中的位置却是最边缘的，也许是由于它只参与了一个事件。行为人 2 和 4 在 2 个维度中都处于边缘，反映他们只参与了一个事件。同样，事件 1 和事件 3 在 2 个维度中也是位于边缘，表明它们比事件 2 吸引的行为人要少一些。

表 5.5 隶属网络的对应分析主坐标

维度(w)	1	2
	行为人/行值(u_{ik})	
A1	−0.661	−0.144
A2	−0.323	0.866
A3	0.000	−1.155
A4	1.323	0.866
A5	0.661	−0.144
	事件/列值(v_{jk})	
E1	−1.146	0.559
E2	0	−0.745
E3	1.146	0.559
奇异值(λ_k)	0.866	0.645

结论 在前面有关章节中,我们讨论了一些网络分析的方法(聚类分析、MDS 和相应分析)。这些方法在用于非网络数据的分析时具有明显的优势。例

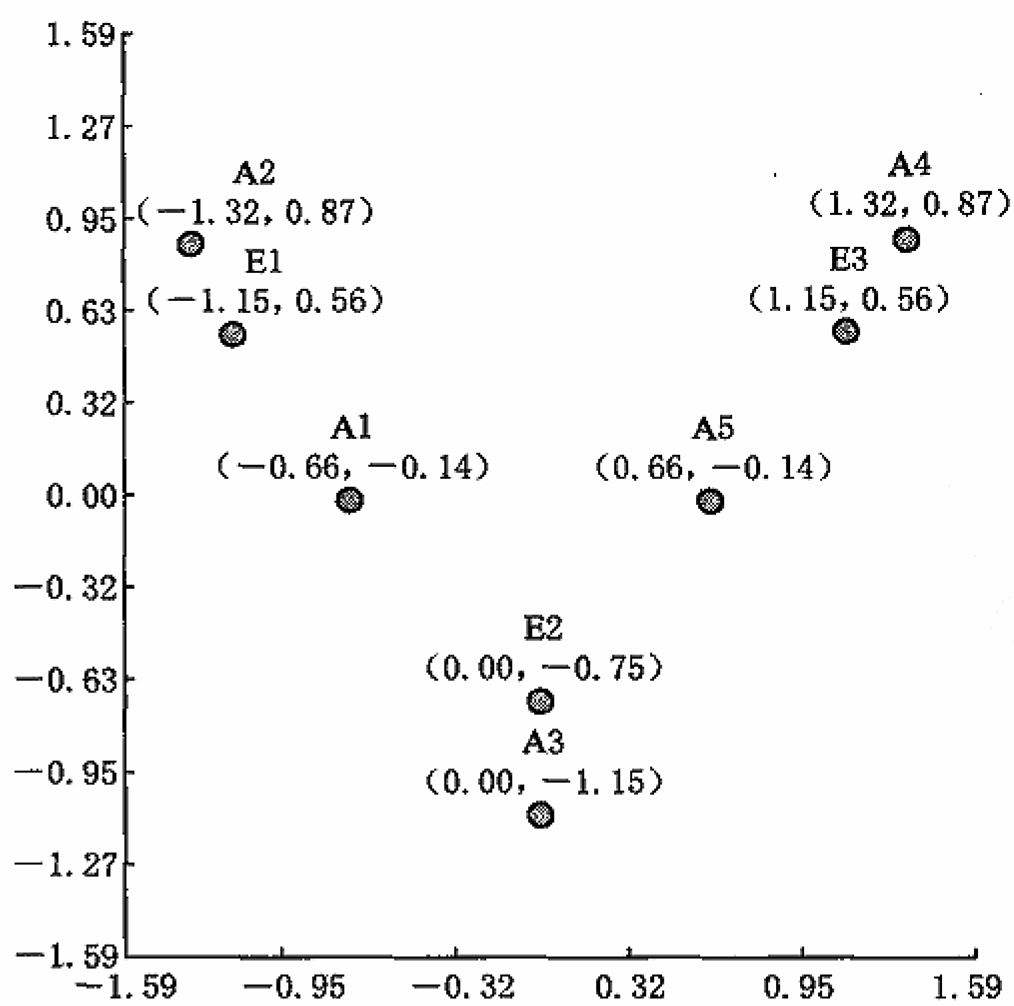


图 5.6 5-行为人和 3-事件隶属网络的对应分析

如，Han(2003)利用加权 MDS 法，分析了 1993 年综合社会调查数据中受访者的音乐爱好。虽然收集社会网络数据耗时费力，尤其是采用全幅网络设计时，但社会调查数据却比较容易获得，可以用这些网络分析方法对它进行分析，以揭示潜在的关系。社会网络分析为研究者提供了有力的分析概念和工具方法。最近有一项研究非常娴熟地运用了聚类分析方法，将中欧东部地区 1586 个市民社会网站根据 13 个变量分类成 6 个维度(Vedres, Bruszt & Stark, 2004)。另一项研究(Mohr, 1998)则直接讨论如何将 MDS、聚类和对应分析等方法用于对复杂结构进行简化，对不同文化背景之间的一些深层结构逻辑进行可视化处理。这些研究展现了多种社会网络方法的熟练应用，也正好表达了我们本书所追求的最终目的——鼓励读者朋友努力学习社会网络分析，以洞悉结构关系之谜。

参考文献

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, *25*, 211—230.

Alba, R. D. (1973). A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, *3*, 113—126.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage.

Alderson, A., & Beckfield, J. (2004). Power and position in the world city system. *American Journal of Sociology*, *109*, 811—851.

Alexander, M. C., & Danowski, J. (1990). Analysis of an ancient network: Personal communication and the study of social structure in a past society. *Social Networks*, *12*, 313—335.

Allison, P. (2001). *Missing data*. Thousand Oaks, CA: Sage.

Anderson, C., Wasserman, S., & Crouch, B. (1999). A p* primer: Logit models for social networks. *Social Networks*, *21*, 37—66.

Anderson, J. G., & Jay, S. (1985). Computers and clinical judgement: The role of physician networks. *Social Science and Medicine*, *20*, 969—979.

Bailey, S., & Marsden, P. (1999). Interpretation and interview context: Examining the General Social Survey name generator using cognitive methods. *Social Networks*, *21*, 287—309.

Baker, W., & Faulkner, R. (1993). The social organization of conspiracy. *American Sociological Review*, *58*, 837—860.

Barabási, A.-L. (2002). *Linked: The new science of networks*. Cambridge, MA: Perseus.

Barnes, J. (1954). Class and committees in a Norwegian island parish. *Human Relations*, *7*, 39—58.

Barnes, J. A. (1979). Network analysis: Orienting notion, rigorous technique or substantive field of study. In P. W. Holland & S. Leinhardt (Eds.), *Perspectives on social network analysis* (pp. 403—423). New York: Academic Press.

Batchelder, E. (2002). Comparing three simultaneous measurements of a sociocognitive network. *Social Networks*, *24*, 261—277.

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, *110*, 44—91.

Beauchamp, M. (1965). An improved index of centrality. *Behavioral Science*, *10*, 161—163.

Benson, M. A. (2004). Dyadic hostility and the ties that bind: State-to-state versus state-to-system security and economic relationships. *Journal of Peace Research*, *41*, 659—676.

Berman, F., & Brady, H. (2005). *Final report: NSF SBE-CISE workshop on cyberinfrastructure and the social sciences*. Washington, DC: National Science Foundation.

Bernard, R. H., Johnsen, E., Killworth, P., McCarty, C., Shelley, G., & Robinson, S. (1990). Comparing four different methods for measuring personal social networks. *Social Networks*, *12*, 179—215.

Bernard, R. H., & Killworth, P. (1977). Informant accuracy in social network data Ⅱ. *Human*

Communications Research, 4, 3—18.

Bernard, R. H., Killworth, P., & Sailer, L. (1981). Summary of research on informant accuracy in network data, and on the reverse small world problem. *Connections*, *4*(2):11—25.

Bernard, R. H., Killworth, P., Sailer, L., & Kronenfeld, D. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, *13*, 495—517.

Blau, P. M., Ruan, D., & Ardelt, M. (1991). Interpersonal choice and networks in China. *Social Forces*, *69*, 1037—1062.

Boorman, S., & White, H. (1976). Social structure from multiple networks. Ⅱ. Role structures. *American Journal of Sociology*, *81*, 1384—1446.

Borgatti, S., & Everett, M. (1992). Notions of position in social network analysis. *Sociological Methodology 22*, 1—35.

Borgatti, S., Everett, M., & Freeman, L. (2004). *UCINET VI software for social network analysis*. Natick, MA: Analytic Technologies.

Borgatti, S., Everett, M., & Shirey, P. (1990). LS sets, lambda sets and other cohesive subsets. *Social Networks*, *12*, 337—357.

Borgatti, S., & Molina, J.-L. (2005). Towards ethical guidelines for network research in organizations. *Social Networks*, *27*, 107—117.

Borgatti, S. P. (2003). The key player problem. In R. Breiger, K. Carley, & P. Pattison (Eds.), *Dynamic social network modeling and analysis: Workshop summary and papers* (pp. 241—252). Washington, DC: National Academy of Sciences Press.

Brandes, U., Kenis, P., Raab, J., Schneider, V., & Wagner, D. (1999). Explorations into the visualization of policy networks. *Journal of Theoretical Politics*, *11*, 75—106.

Brewer, D. D. (2000). Forgetting in the recall-based elicitation of personal and social networks. *Social Networks*, *22*, 29—43.

Brewer, D. D., & Webster, C. (1999). Forgetting of friends and its effects on measuring friendship networks. *Social Networks*, *21*, 361—373.

Burgin, R. (1995). The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *Journal of the American Society for Information Science*, *46*(8), 562—572.

Burt, R. (1983). Range. In R. S. Burt & M. Minor(Eds). *Applied network analysis: A methodological introduction*. Beverly Hills, CA: Sage.

Burt, R. (1985). General social survey network items. *Connections*, *8*, 119—123.

Burt R. (1987). A note on missing network data in the General Social Survey. *Social Networks*, *9*, 63—73.

Burt, R. S. (1978). Cohesion versus structural equivalence as a basis for network subgroups. *Sociological Methods and Research*, *7*, 189—212.

Burt, R. S. (1979). Disaggregating the effect on profits in manufacturing industries of having imperfectly competitive consumers and suppliers. *Social Science Research*, *8*(2), 120—143.

Burt, R. S. (1983). Distinguishing relational contents. In R. S. Burt & M. J. Minor (Eds.), *Applied network analysis* (pp. 35—74). Beverly Hills, CA: Sage.

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Har-

vard University Press.

Campbell K. E. , & Lee, B. (1991). Name generators in surveys of personal networks. *Social Networks*, *13*, 203—221.

Campbell, K. E. , Marsden, P. , & Hurlbert, J. (1986). Social resources and socioeconomic status. *Social Networks*, *8*, 97—117.

Carley, K. , & Krackhardt, D. (1996). Cognitive inconsistencies and non-symmetric friendship. *Social Networks*, *18*, 1—27.

Carrington, P. , Scott, J. , & Wasserman, S. (Eds.). (2005). *Models and methods in social network analysis*. New York: Cambridge University Press.

Casciaro, T. (1998). Seeing things clearly: Social structure, personality, and accuracy in social network perception. *Social Networks*, *20*, 331—351.

Cohen, S. , Frank, E. , Doyle, W. J. , Skoner, D. P. , Rabin, B. S. , & Gwaltney, J. M. , Jr. (1998). Types of stressors that increase susceptibility to the common cold in adults. *Health Psychology*, *17*, 214—223.

Coleman, J. S. (1986). Social theory, social research, and a theory of action. *American Journal of Sociology*, *91*, 1309—1335.

Cowan, R. , & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, *28*, 1557—1575.

Cross, R. , Liedtka, J. , & Weiss, L. (2005). A practical guide to social networks. *Harvard Business Review*, *83*(3), 124—132.

Crouch, B. , & Wasserman, S. (1998). A practical guide to fitting social network models via logistic regression. *Connections*, *21*, 87—101.

Davis, J. A. (1979). The Davis/Holland/Leinhardt studies: An overview. In P. W. Holland & S. Leinhardt(Eds.). *Perspectives on social network research*(pp. 51—62). New York: Academic Press.

Degenne, A. , & Forse, M. (1990). *Introducing social networks*. Thousand Oaks, CA: Sage.

Diani, M. , & McAdam, D. (Eds.). (2003). *Social movements and networks: Relational approaches to collective action*. New York: Oxford University Press.

Doreian, P. , Batagelj, V. , & Ferligoj, A. (2005). *Generalized blockmodeling*. Cambridge. UK: Cambridge University Press.

Doreian, P. , & Woodard, K. (1992). Fixed list versus snowball selection of social networks. *Social Science Research*, *21*, 216—233.

Doreian, P. , & Woodard, K. (1994). Defining and locating cores and boundaries of social networks. *Social Networks*, *16*, 267—293.

Dunbar, R. I. M. , & Spoor, M. (1995). Social networks, support cliques and kinship. *Human Nature*, *6*, 273—290.

Duquenne, V. (1996). On lattice approximations: Syntactic aspects. *Social Network*, *18*, 189—199.

Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology*, *103*, 281—317.

Erickson, B. H. (1996). Culture, class, and connections. *American Journal of Sociology*, *102*, 217—251.

Erickson, B. H. (2004). The distribution of gendered social capital in Canada. In H. Flap & B. Volker(Eds.), *Creation and returns of social capital: A new research program*(pp. 27—51). New York: Routledge.

Everett, M. (1985). Role similarity and complexity in social networks. *Social Networks*, *7*, 353—359.

Everett, M., Boyd, J., & Borgatti, S. (1990). Ego-centreed and local roles: A graph theoretic approach. *Journal of Mathematical Sociology*, *15*, 163—172.

Fafchamps, M., van der Leij, M. J., & Goyal, S. (2006). *Scientific networks and co-authorship (No. 256)*. University of Oxford, Department of Economics, Economics Series Working Papers.

Faust, K. (1988). Comparison of methods for positional analysis: Structural and general equivalences. *Social Networks 10*, 313—341.

Faust, K. (1997). Centrality in affiliation networks. *Social Networks 19*, 157—191.

Faust, K. (2005). Using correspondence analysis for joint displays of affiliation networks. In P. J. Carrington, J. Scott, & S. Wasserman(Eds.), *Models and methods in social network analysis* (pp. 117—147). New York: Cambridge University Press.

Faust, K., & Skvoretz, J. (2002). Comparing networks across space and time, size and species. *Sociological Methodology*, *32*, 267—299.

Feld, S. L., & Carter, W. C. (2002). Detecting measurement bias in respondent reports of personal networks. *Social Networks*, *24*, 365—383.

Feldman-Savelsberg, P., Ndonko, F. L., & Yang, S. (2005). How rumor begets rumor: Collective memory, ethnic conflict, and reproductive rumors in Cameroon. In G. A. Fine, V. Campion-Vincent, & C. Heath(Eds), *Rumor mills: The social impact of rumor and legend* (pp. 141—159). New York: Transaction Press.

Fischer, C. S. (1982). What do we mean by "friend"? An inductive study. *Social Networks 3*, 287—306.

Fischer, C. S., & Shavit, Y. (1995). National differences in network density: Israel and the United States. *Social Networks 17*, 129—145.

Frank, O. (2003). *KliqFinder for Windows, version 0.05*. East Lansing: Michigan State University.

Frank, O. (2005). Network sampling and model fitting. In P. Carrington, J. Scott, & S. Wasserman(Eds.), *Models and methods in social network analysis* (pp. 31—45). New York: Cambridge University Press.

Frank, O., & Snijders, T. A. B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, *10*, 53—67.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association 81*, 832—842.

Freeman, L. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, *40*,

35—41.

Freeman, L. (1979). Centrality in social networks: I. Conceptual clarification. *Social Networks*, *1*, 215—239.

Freeman, L. C. (1992). Filling in the blanks: A theory of cognitive categories and the structure of social affiliation. *Social Psychology Quarterly*, *55*, 118—127.

Freeman, L. C. (1992). The resurrection of cliques: Application of Galois lattices. *Bulletin de Methodologie Sociologique 37*, 3—24.

Freeman, L. C. (1996). Some antecedents of social network analysis. *Connections*, *19*, 39—42.

Freeman, L. C. (2000). Visualizing social networks, *Journal of Social Structure*, *1*, 1—15.

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, Canada: Empirical Press.

Freeman, L. C. (2005). Graphic techniques for exploring social network data. In P. J. Carrington, J. Scott, & S. Wasserman(Eds.), *Models and methods in social network analysis* (pp. 248—270). Cambridge, MA: Cambridge University Press.

Freeman, L. C., Borgatti, S., & White, D. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, *13*, 141—154.

Freeman, L. C., Romney, K., & Freeman, S. (1987). Cognitive structure and informant accuracy. *American Anthropologist*, *89*(2), 310—325.

Freeman, L. C., & Thompson, C. R. (1989). Estimating acquaintanceship volume. In M. Kochen (Ed.), *The small world* (pp. 147—158). Norwood, NJ: Ablex.

Freeman, L. C., & Webster, C. (1994). Interpersonal proximity in social and cognitive space. *Social Cognition*, *12*(3), 223—247.

Freeman, L. C., & White, D. (1993). Using Galois lattices to represent network data. *Sociological Methodology*, 23, 127—146.

Galaskiewicz, J. (1979). The structure of community organizational networks. *Social Forces*, *57*, 1346—1364.

Goodman, L. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, *32*, 148—170.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, *78*, 1360—1380.

Han, S. (2003). Unraveling the brow: What and how of choice in musical preference. *Sociological Perspectives*, *46*, 435—459.

Harary, F. (1969). *Graph theory*. Reading, MA: Addison-Wesley.

Hargens, L. L. (2000). Using the literature: Reference networks, reference contexts, and the social structure of scholarship. *American Sociological Review*, *65*, 846—865.

Harrisson, D., Laplante, N., & St-Cyr, L. (2001). Cooperation and resistance in work innovation networks. *Human Relations*, *54*, 215—255.

Heckathorn, D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, *44*, 174—199.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for di-

rected graphs. *Journal of the American Statistical Association*, *76*, 33—65.

Huisman, M., & van Duijn, M. A. J. (2005). Software for social network analysis. In P. J. Carrington, J. Scott, & S. Wasserman(Eds.), *Models and methods in social network analysis* (pp. 270—316). New York: Cambridge University Press.

Johnson, J., & Orbach, M. (2002). Perceiving the political landscape: Ego biases in cognitive political networks. *Social Networks*, *24*, 291—310.

Jolly, A. M., Muth, S. Q., Wylie, J. L., & Potterat, J. J. (2001). Sexual networks and sexually transmitted infections: A tale of two cities. *Journal of Urban Health*, *78*(3), 433—445.

Jones, L. M., & Fischer, C. S. (1978). *Studying egocentric network by mass survey*. Working Paper at the Institute of Urban and Regional Development, University of California, Berkeley.

Kang, S.-Y., Deren, S., Andia, J., Colon, H. M., & Robles, R. (2005). Egocentric HIV risk networks among Puerto Rican crack users in New York and in Puerto Rico: Impact on sex risk behaviors over time. *AIDS Education and Prevention*, *17*, 53—67.

Kenis, P., & Knoke, D. (2002). How organizational field networks shape interorganizational tie-formation rates. *Academy of Management Review*, *27*, 275—293.

Killworth, P., Johnsen, E., Bernard, R., Shelley, G., & McCarty, C. (1990). Estimating the size of personal networks. *Social Networks*, *12*, 289—312.

Killworth, P. D., Bernard, R., & McCarty, C. (1984). Measuring patterns of acquaintanceship. *Current Anthropology*, *25*, 381—392.

Kirke, D. M. (1996). Collecting peer data and delineating peer networks in a complete network. *Social Networks*, *18*, 333—346.

Klovdahl, A. S. (2005). Social network research and human subjects protection: Towards more effective infectious disease control. *Social Networks*, *27*, 119—137.

Knoke, D. (2001). *Changing organizations: Business networks in the new political economy*. Boulder, CO: Westview Press.

Knoke, D. (in press). Playing well together: Creating corporate social capital in strategic alliance networks. *American Behavioral Scientist*.

Knoke, D., & Burt, R. (1983). Prominence. In R. S. Burt & M. J. Miner(Eds.), *Applied network analysis: A methodological introduction*(pp. 195—222). Beverly Hills CA: Sage.

Knoke, D., & Kuklinski, J. H. (1982). *Network analysis*. Beverly Hills, CA: Sage.

Knoke, D., & Laumann, E. O. (1982). The social structure of national policy domains: An exploration of some structural hypotheses. In P. V. Marsden & N. Lin(Eds.), *Social structure and network analysis*(pp. 255—270). Beverly Hills, CA: Sage.

Knoke, D., Pappi, F. U., Broadbent, J., & Tsujinaka, Y. (1996). *Comparing policy networks: Labor politics in the U. S., Germany, and Japan*. New York: Cambridge University Press.

Knoke, D., & Rogers, D. (1979). A blockmodel analysis of interorganizational networks. *Sociology and Social Research*, *64*, 28—52.

Knox, H., Savage, M., & Harvey, P. (2006). Social networks and the study of relations: Networks as method, metaphor and form. *Economy and Society*, *35*, 113—140.

Koehly, L. M., & Pattison, P. (2005). Random graph models for social networks: Multiple rela-

tions or multiple raters. In P. J. Carrington, J. Scott, & S. Wasserman(Eds.), *Models and methods in social network analysis*(pp. 162—191). New York: Cambridge University Press.

Kossinets, G. (2003). Effects of missing data in social networks. *Social Networks*, *28*, 247—268.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, *9*, 109—134.

Krackhardt, D., & Kilduff, M. (1999). Whether close or far: Perceptions of balance in friendship networks in organizations. *Journal of Personality and Social Psychology*, *76*, 770—782.

Kronenfeld, D. B., & Kronenfeld, J. (1972). Toward a science of design for successful food service. *Institutions and Volume Feeding Management*, *70*, 38—44.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Kumbasar, E., Romney, K., & Batchelder, W. (1994). Systematic biases in social perception. *American Journal of Sociology*, *100*, 477—505.

LaPierre, R. T. (1934). Attitudes vs. actions. *Social Forces*, *13*, 230—237.

Laumann, E., Marsden, P. V., & Prensky, D. (1983). The boundary-specification problem in network analysis. In R. Burt & M. Minor(Eds.), *Applied network analysis* (pp. 18—34). Beverly Hills, CA: Sage.

Laumann, E. O., & Knoke, D. (1987). *The organizational state: Social choice in national policy domains*. Madison: University of Wisconsin Press.

Laumann, E. O., Mardsen, P. V., & Prensky, D. (1989). The boundary specification problem in network analysis. In L. C. Freeman, D. R. White, & A. K. Romney(Eds.), *Research methods in social network analysis*(pp. 61—87). Fairfax, VA: George Mason University Press.

Lin, N. (1982). Social resources and instrumental action. In P. V. Marsden & N. Lin(Eds.), *Social structure and network analysis*(pp. 131—145). Beverly Hills, CA: Sage.

Lin, N., & Dumin, M. (1986). Access to occupations through social ties. *Social Networks*, *8*, 365—385.

Lin, N., Fu, Y., & Hsung, R.-M. (2001). The position generator: Measurement techniques for investigations of social capital. In N. Lin, K. Cook, & R. Burt(Eds.), *Social capital: Theory and research* (pp. 57—81). Hawthorne, NY: Aldine de Gruyter.

Luce, D., & Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika*, *14*, 95—116.

Marsden, P. (2002). Egocentric and sociocentric measures of network centrality. *Social Networks*, *24*, 407—422.

Marsden, P. V. (1987). Core discussion networks of Americans. *American Sociological Review*, *52*, 122—131.

Marsden, P. V. (1993). The reliability of network density and composition measures. *Social Networks*, *15*(4), 399—421.

Marsden, P. V. (2005). Recent developments in network measurement. In P. Carrington, J. Scott, & S. Wasserman, *Models and methods in social network analysis* (pp. 8—30). New York: Cambridge University Press.

McCarty, C., Bernard, R., Killworth, P., Shelley, G., & Johnsen, E. (1997). Eliciting representative samples of personal networks. *Social Networks*, *19*, 303—323.

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in America: Changes in core discussion networks over two decades. *American Sociological Review*, *71*, 353—375.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, *49*(1), 93—123.

Mische, A., & Pattison, P. E. (2000). Composing a civic arena: Publics, projects, and social settings. *Poetics*, *27*, 163—194.

Mitchell, J. C. (1969). *Social networks in urban situations: Analyses of personal relationships in Central African towns*. Manchester, UK: Manchester University Press.

Mohr, J. W. (1998). Measuring meaning structures. *Annual Review of Sociology*, *24*, 345—370.

Mohr, J. W., & Duquenne, V. (1997). The duality of culture and practice: Poverty relief in New York City, 1888—1917. *Theory and Society*, *26*, 305—356.

Mokken, R. J. (1979). Cliques, clubs and clans. *Quantity and Quality*, *13*, 161—173.

Moreno, J. L. (1934). *Who shall survive?* Washington, DC: Nervous and Mental Disease Publishing Company.

Morrissey, J., Tausig, M., & Lindsey, M. (1985). *Network analytic methods for mental health service system research: A comparison of two community support systems (No. 6)*. Washington, DC: National Institute of Mental Health Series BN.

Mouttapa, M., Valente, T. W., & Gallaher, P. (2004). Social network predictors of bullying and victimization. *Adolescence*, *39*, 315—335.

Nadel, S. F. (1957). *The theory of social structure*. London: Cohen and West.

Neuman, W. L. (2000). *Social research methods: Qualitative and quantitative approaches* (4th ed.). Boston: Allyn & Bacon.

Newcomb, T. (1961). *The acquaintance process*. New York: Holt, Rinehart and Winston.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167—256.

de Nooy, W., Mrvar, W. A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. New York: Cambridge University Press.

Nowicki, K. & Snijders T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077—1088.

Nyblom, J., Borgatti, S., Roslakka, J., & Salo, M. A. (2003). Statistical analysis of network data: An application to diffusion of innovation. *Social Networks*, *25*, 175—195.

Padgett, J. F. & Ansell, C. K. (1993). Robust action and the rise of the Medici, 1400—1434. *American Journal of Sociology*, *98*, 1259—1319.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

Pattison, P. (1988). Network models: Some comments on papers in this special issue. *Social Networks*, *10*, 383—411.

Pattison, P., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: Ⅱ. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, *52*, 169—193.

Peay, E. R. (1980). Connectedness in a general model for valued networks. *Social Networks*, *2*, 385—410.

Perrow, C. (1986). *Complex organizations: A critical essay*. New York: Random House.

Podolny, J. M. & Stuart, T. (1995). A role-based ecology of technological change. *American Journal of Sociology*, *100*, 1224—1260.

Raab, J., & Milward, H. B. (2003). Dark networks as problems. *Journal of Public Administration Research and Theory*, *13*, 413—439.

Rice, R. (1990). Computer-mediated communication system network data: Theoretical concerns and empirical examples. *International Journal of Man-Machine Studies*, *32*, 627—647.

Rice, R. (1994). Network analysis and computer-mediated communication systems. In S. Wasserman & J. Galaskiewicz(Eds.), *Advances in social network analysis* (pp. 167—203). Newbury Park, CA: Sage.

Rice, R., Borgman, C., Bednarski, D., & Hart, P. J. (1989). Journal-to-journal citation data: Issues of reliability and validity. *Scientometrics*, *15*, 257—282.

Robins, G., Pattison, P., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: Ⅲ. Valued relations. *Psychometrika*, *64*, 371—394.

Romney, K., & Weller, S. (1984). Predicting informant accuracy from patterns of recall among individuals. *Social Networks*, *6*, 59—77.

Romney, K., Weller, S., & Batchelder, W. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*, 313—338.

Ruan, D. (1998). The content of the General Social Survey discussion networks: An exploration of General Social Survey discussion name generator in a Chinese context. *Social Networks*, *20*, 247—264.

Ruan, D., Freeman, L., Dai, X., Pan, Y., & Zhang, W. (1997). On the changing structure of social networks in urban China. *Social Networks*, *19*, 75—89.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*, 581—603.

Salganik, M. J., & Heckathorn D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, *34*, 193—240.

Schiffman, S., Reynolds, L., & Young, F. (1981). *Introduction to multidimensional scaling*. New York: Academic Press.

Schwartz, J. E. (1977). An examination of CONCOR and related methods for blocking sociometric data. *Sociological Methodology*, *7*, 255—282.

Schweizer, T. (1996). Actor and event orderings across time: Lattice representation and Boolean analysis of political disputes in Chen Village, China. *Social Networks*, *18*, 247—266.

Scott, J. (1991). *Social network analysis: A handbook*. London: Sage.

Scott, J. (2000). *Social network analysis: A handbook* (2nd ed.). Thousand Oaks, CA: Sage.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, *5*, 269—287.

Simmel, G. (1908). *Sociology: Investigations on the forms of sociation*. Leipzig, Germany: Duncker and Humblot.

Slez, A., & Martin, J. L. (2007). Political action and party formation in the United States consti-

tutional convention. *American Sociological Review*, *72*(1), 42—67.

Smith, D., & Timberlake, M. (2002). Hierarchies of dominance among world cities: A network approach. In S. Sassen(Ed.), *Global networks, linked cities* (pp. 117—141). New York: Routledge.

Snijders, T. A. B., Steglich, C., Schweinberger, M., & Huisman, M. (2007). *Manual for SIENA version 3*. Groningen, Netherlands: University of Groningen, ICS, Department of Sociology.

Stoebenau, K., & Valente, T. W. (2003). Using network analysis to understand community-based programs: A case study from highland Madagascar. *International Family Planning Perspectives*, *29*(4), 167—173.

Stork, D., & Richards, W. (1992). Nonrespondents in communication network studies: Problems and possibilities. *Group & Organization Management*, *17*(2), 193—210.

Strang, G. (1988). *Linear algebra and its applications* (3rd ed.). San Diego: Harcourt, Brace, Jovanovich.

Strauss, D., & Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, *85*, 204—212.

Taylor, S., & Fiske, S. (1978). Salience, attention, and attribution: Top-of-the-head phenomena. In L. Berkowitz(Ed.), *Advances in experimental social psychology*, (*vol. 11*) pp. 249—288. New York: Academic Press.

Tenkasi, R. V., & Chesmore, M. C. (2003). Social networks and planned organizational change: The impact of strong network ties on effective change implementation and use. *Journal of Applied Behavioral Science*, *39*, 281—300.

Tsuji, R. (1997). *PermNet, version 0.94*. Sapporo: Hokkaido University.

Tsvetovat, M., & Carley K. M. (2005). Structural knowledge and success of anti-terrorist activity: The downside of structural equivalence. [Electronic version]. *Journal of Social Structure*, *6*(2).

Useem, M. (1979). The social organization of the American business elite and participation of corporation directors in the governance of American institutions. *American Sociological Review*, *44*, 553—572.

Valente, T. (1995). *Network models of the diffusion of innovations*. Cresskill, NJ: Hampton Press.

Van der Gaag, M., & Snijders, T. (2004). Proposals for the measurement of individual social capital. In H. Flap & B. Volker(Eds.), *Creation and returns of social capital: A new research program*(pp. 199—219). New York: Routledge.

Vedres, B., Bruszt, L., & Stark, D. (2004). Organizing technologies: Genre forms of online civic association in Eastern Europe. *Annals of the American Academy of Political and Social Science*, *597*, 171—188.

Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regressions for social networks:

Ⅰ. An introduction to Markov graphs and p. *Psychometrika*, *61*(3), 401—425.

Wasserman, S. , & Robins, G. (2005). An introduction to random graphs, dependence graphs, and p*. In P. J. Carrington, J. Scott, & S. Wasserman(Eds.), *Models and methods in social network analysis*(pp. 148—161). New York: Cambridge University Press.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. New York: Norton.

Wellman, B. (1999). From little boxes to loosely bounded networks: The privatization and domestication of community. In J. L. Abu-Lughod(Ed.), *Sociology for the twenty-first century* (pp. 94—114). Chicago: University of Chicago Press.

Wellman, B. , & Berkowitz, S. D. (1988). *Social structures: A network approach*. Cambridge UK: Cambridge University Press.

White, D. (1996). Statistical entailments and the Galois lattice. *Social Networks*, *18*, 201—215.

White, D. , & Reitz, K. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks*, *5*, 193—234.

White, H. C. , Boorman, S. A. , & Breiger, R. L. (1976). Social structure from multiple networks. I: Blockmodels of roles and positions. *American Journal of Sociology*, *81*, 730—780.

White, K. , & Watkins, S. C. (2000). Accuracy, stability and reciprocity in informal conversational networks in rural Kenya. *Social Networks*, *22*, 337—355.

White, M. , Kim, A. , & Glick, J. J. (2005). Mapping social segregation. *Sociological Methods and Research*, *34*, 173—203.

Winship, C. , & Mandel, M. (1983). Roles and positions: A critique and extension of the blockmodeling approach. *Sociological Methodology*, *14*, 314—344.

Wu, L. L. (1983). Local blockmodel algebras for analyzing social networks. *Sociological Methodology*, *14*, 272—313.

Yang, S. , & Hexmoor, H. (2004). Measuring optimal connections in large networks: A new algorithm and its application. *Journal of Mathematical Sociology*, *28*(3), 197—213.

Yang, S. , & Knoke, D. (2001). Optimal connections: Strengths and distance in valued graphs. *Social Networks*, *23*, 285—295.

译名对照表

actors	行为人
alters	行为客体
actor contacts	行为人联系
adjacency matrix	邻接矩阵
advice-seeking network	建议寻求网络
affiliation matrix	隶属矩阵
affiliation networks	隶属网络
Americans' core discussion groups	美国人的核心讨论圈
analysis of lattices	点阵分析
archival documents	档案资料
a single inclusive cluster	单一的全涵盖聚类
asymmetric choice	非对称选择
attributes	特征
authority/power relations	权力关系
automorphic equivalence	自形等价
average link	平均联结
average method	均值法
betweenness	中间性
betweenness centrality	中间性中心度
binary nondirected	二进制无向
binary directed	二进制有向
bipartite graph	二分图
blockmodel	块模型
boundary penetration relations	交叉关系
boundary specification	范围界定
bridge	桥
cascade modules	级联模块
categoric variables	分类变量
centrality	中心度
chain-referral method	连锁推荐法
chaos theory	混沌理论
checklist	提名清单
checklist method	名单核对法
circuit	回路
clique	小团体
clique size	小团体规模

closed walks	封闭行程
closeness	亲密性
closeness Centrality	亲密中心度
cluster	群
clustering	聚类
clustering algorithms	聚类演绎法
cognitive maps	认知图
cognitive social structure(CSS)	认知社会结构
cohesion	内聚力
cohesiveness	内聚
cohesive subgroup	内聚型亚群体
communication relations	沟通关系
complementarity hypothesis	异性相吸的互补假设
complete link	完全联结
complete reliability	完全可信
complete networks	全幅网络
complete network measures	全幅网络测量
complex theory	复杂理论
components	组成部分
computer-mediated communication(CMC)	电子化交流
computer-monitored network data	计算机控制网络数据
connection	连接
consensus formula	共识模型
contact	接触
correspondence analysis	对应分析
cutpoints	分界点
degree centrality	程度中心度
dendrogram	树形图
density cutoff	密度临界值
density matrix	密度矩阵
diameters	直径
digraph	有向图
digraph clique	有向图小团体
directed dyadic relations	有向对偶关系
directed graph	有向图表
direct relation	定向关系
dissimilarities	异质性
diversity	多样性
divisiveness	分裂性
descriptive statistics	描述统计学
dispersion	离差

drawn sociograms	手绘社会关系图
dyad	对点
dyadic network	对偶网络
dyadic relation	对偶网络关系
dyadic units	对偶组合
ego	行为主体
ego actor	主体行为人
egocentric analysis	自我中心分析法
egocentric networks	自我中心网络
egocentric network measures	自我中心网络的测量
ego-density	自我密度
eigenvector centrality	特征向量中心度
error term	误差项
Euclidean distance	欧氏距离
event	事件
event-based strategies	基于事件的策略
event contacts	事件联系
expansiveness	扩展程度
faction	派系
first zone	第一关系圈
fixed list approach	固定名单法
fixed list selection	固定名单选择法
frequency	频率
galois connection	伽罗瓦连接
gang	帮派
general social survey(GSS)	综合社会调查
geodesic distance	测地距
geographical constraint	地域限制
graphs	图表法
graph theory	图论
group closeness centralization	组亲密中心度
group degree centralization	组程度中心度
hierarchical agglomerative cluster analysis	分层聚类分析
hierarchical clustering	分层聚类法
high degree-seeking strategy	高相关度寻找法
homophily hypothesis	羽禽群居的同质假设
human capital	人力资本
indegrees	节点入度
index of qualitative variation(IQV)	定性变化指数
individual differences scaling(INDSCAL)	个体差异标度

informant bias	信息偏差
initial sample	初始样本
initiators	发起人
instrumental name generator	工具性提名生成法
instrumental relations	利用关系
instrumental ties	工具性连接点
intensity	强度
interlocking roles	互锁角色
internal consistency reliability	内在一致性信度
interobserver reliability	观察者间信度
isolated structures	孤立结构
isomorphism	同质性
isomorphic equivalence	同构等价
isomorphic subgraphs	同构子图
Jaccard's coefficient	Jaccard 系数
jointly dyadic property	共同的对偶特征
K-core method	K 核心法
Kevin Bacon game	凯文·培根游戏
kinship and descent relations	亲属关系
labeled points	标示点
lattices	点阵
logit	模型 p*
longitudinal information	纵向信息
matrices	矩阵法
maximal complete subgraph	最大完全子图
mean egocentric density measure	自我中心网络密度测量均值
mean sex diversity	性别多样化均值
measurement of relations	关系测量
missing data	数据缺失
multidimensional scaling(MDS)	多维标度
multiple-name generator methods	多个提名生成法
multiple networks	多重网络
multiple social support name generator	多重社会资源提名法
mutualchoice	双向选择
name generators	提名法
name interpreters	姓名阐释法
national Longitudinal Study of Adolescent Health	美国青少年健康纵向调查
network position measures	网络位置测量
network sampling	网络抽样
network size	网络规模

network structure	网络结构
nodes	节点
nodal degree	节点度
nodal indegree	节点入度
nodal outdegree	节点出度
nodal response rate	节点应答率
nominalist strategies	唯名论界定策略
nondirected graph	非定向图
nondirected dyadic relations	非定向的对偶关系
nonoverlapping clusters	非交叉聚类
nonreciprocal relations	单向关系
normalized actor degree centrality	标准化行为人程度中心度
numerical constraint	数量限制
one-mode network	一模网络
online social networking	在线网络交流
ordinary least square regression methods	最小二乘回归法
pair characteristics	二元特性
paths	路径
Pearson's correlation coefficient	皮尔逊相关系数
points	点
popularity	受欢迎程度
positon	位置
positional generators	位置选择法
positional strategies	定位策略
prestige	声望
realist strategies	唯实论界定策略
receivers	接收者
receptivity	接纳程度
regular equivalence	正规等价
relations	关系
relational response rate	关系应答率
relational strategies	关系策略
relationship measures	关系的测量方法
reliability	信度
reputational method	声望法
resource generators	资源选择法
respondent-driven sampling	应答驱动抽样
reverse small-world method(RSW)	反向小世界方法
role/content constraint	角色/内容限制

sampling bias	抽样偏差
sampling frame	抽样框
sampling units	抽样单位
saturated structures	饱和结构
seciometry	社会计量法
self-relation	自我关系
self-report questionaire surveys	自填式问卷调查
sender-receiver	发送者—接收者
sentiment relations	情感关系
similarity	相似性
single link	单个联结
single-name generator methods	单个提名生成法
singleton cluster	单独的聚类
singular value decomposition(SVD)	奇异值分解
six degree of separation	六度分隔
small-world idea	小世界理论
small-world phenomena	小世界现象
snowball sampling	滚雪球抽样法
social capital	社会资本
social cognitive structure	认知社会结构
social identity approaches	社会认同途径
social resources theory	社会资源理论
social support networks	社会支持网络
sociogram	社会关系图
sociomatrix	社会矩阵
sociometric	社会计量学
sociometric techniques	社会计量方法
source-target	源头—目标
split-half reliability	折半信度
standard deviation	标准差
standard deviation for continuous variables	连续变量的标准差
star graph	星形图
strength	力度
stress	应力
strongly connected	强连接
structural equivalence	结构等价
structural relations	结构关系
structurally equivalent actors	结构等价的行为人
subgraph	子图
subgroup	亚群体

substantialist approaches	实体论方法
subsystem	子系统
summation method	求和法
temporal constraint	时间限制
terminators	终结者
test-retest reliability	再测法信度
total network	全部网络
transaction relations	交换关系
tree diagram	树图
triadic relations	三元关系
two-mode network	二模网络
type of tie	连接模式
unilaterally connected	单向连接
utility	效用
valued directed graph	赋值有向图
valued graph	赋值图
valued nondirected graph	赋值无向图
visual displays	视觉演示
weak connected	弱连接
weighted multidimensional scaling(WMDS)	加权多维标度

因子分析：统计方法与应用问题

作 者 简 介

金在温(JAE-ON KIM)

美国爱荷华大学社会学系教授,亚太研究中心主任,于加利福尼亚大学伯克利分校获得博士学位。专注于研究政治社会学、社会不平等、定量方法和东亚社会。在《美国社会学刊》(*American Journal of Sociology*),《美国政治科学评论》(*American Political Science Review*),《美国社会学评论》(*American Sociological Review*)和《社会力》(*Social Forces*)等杂志上发表过文章,与维巴(Verba)、聂(Nie)合著《*Equality and Political Participation: A Seven Nation Comparison*》(1978)一书。现在主要研究:三个东亚国家——中国、日本和韩国的价值比较;对政党形成和选举制度的自然实验的研究;为东亚社会建立综合社会调查的合作项目;关于在社会科学研究中处理不完美数据和不完整理论等问题的敏感度分析。

查尔斯·W. 米勒(Charles W. Mueller)

美国爱荷华大学社会学系教授。在《美国社会学评论》(*American Sociological Review*),《社会学方法和研究》(*Sociological Methods and Research*)和《工作与职业》(*Work and Occupations*)等杂志上发表过社会分层和定量方法的文章。现在主要从事组织改组、雇员忠诚度和满意度方面的研究。

译 者 简 介

叶华

曾获中山大学社会学学士学位和硕士学位,现为香港科技大学社会科学部博士候选人,研究兴趣为社会分层与流动、经济社会学、教育社会学。

序

本书扩展了 Jae-On Kim 和 Charles W. Mueller 的专著《因子分析导论:它是什么以及如何运用》(*Introduction to factor analysis: what it is and how to do it*)。后者着重对因子分析技术基础的介绍,关心的是"为什么进行因子分析"这个问题,并且希望读者询问运用这种方法涉及哪些假定。

对阅读使用因子分析方法的书或文章,或者用自己的数据来简单试用因子分析方法的读者来说,本书将把他们引导到另一个阶段。在书中,Kim 和 Mueller 用更为明确的数据分析例子,更为详细地介绍了因子分析的不同方法,以及它们在何种情况下最有用。验证性和探索性因子分析的差别在这里会比在《因子分析导论:它是什么以及如何运用》中讨论得更深入,同样会进行讨论的还有因子旋转的各种标准。特别值得一提的是对不同形式的斜交旋转的讨论,以及如何解释从这些分析中得到的各项系数。Kim 和 Mueller 也回答了在探索性因子分析中抽取出的因子数量的问题,讨论了在验证性因子分析中检验假设的方法。对于这些研究中抽取出来的尺度得分,他们也讨论了分析它们时会遇到的问题。本书会提供一份译名对照表,同时还会解答使用因子分析技术的人经常提出的一些问题。读者可能会被其中一些问题的回答所困扰,但只要加以注意,分析者就能避免错误的跳跃性推论。本书尽管有一些矩阵代数的应用,但其数理内容基本上是自明的。如果读者不能很好地理解矩阵应用,建议参考作者之前的那本专著。

在经济学中,因子分析被用于抽取一组不相关的变量来做进一步分析,因为此时使用高度相关的变量在回归分析中可能会产生误导性的结果。政治学家比较不同国家在一系列政治和社会经济变量上的特点,尝试找出什么是区分国家的最重要特征(例如富裕程度和国家大小)①;社会学家通过研究相互联系最多的人(而不是与其他人联系),来确定"朋友群体"。心理学家和教育学者使用这种技术,研究人们对不同的"激励"有什么感受,并把它们区分为不同的响应组,例如语言的不同要素是相互联系的。

---

① 参见 Rudolph J. Rummel, *The Dimensions of Nations* (Beverly Hills. CA: Sage Publications, 1972)。

正如作者所指出的，本书不可能涵盖因子分析的所有方面，因为在这个领域里经常会有新的发展。然而，如果读者能够建立起系统性的知识，知道如何运用这种技术，以及我们实际上至少在无形中做了什么假定，那么本书就已经达到其目的了。

——E.M.尤斯拉纳(E.M.Uslaner)

第 1 章 | 导论

因子分析概念性的基础简单易学。然而，一些原因可能使得在实际运用中掌握这种方法会比较困难。首先，一般而言，理解统计估计的原则相比理解背后的概念模型，前者对数理复杂性的要求更高。第二，文献中提出了许多提取因子的方法，即使是相对简单的计算机程序，也可能在分析的每一步给出许多选项。这些复杂性会让初学者很难适应，即使专家也会感到局促。第三，我们平时遇到的实际研究问题几乎总是比因子分析模型所假定的情况要复杂。例如，(1)可能会有一些或者全部变量的测量层次都没有达到因子分析的要求；(2)对某些人所用的数据来说，模型的某些假定是不现实的，例如测量误差之间不相关；(3)我们可能会有一些次要的因子，对它们的识别并不是我们考虑的重点，但是它们的存在会影响我们对主要的共同因子的识别。问题的关键在于，研究者最终必须自行作出一些与统计无关的决定。幸运的是，正如我们将会讲述的那样，这些困难是可以克服的。

正如我们在《因子分析导论：它是什么以及如何运用》所指出的，研究者或多或少地被迫依赖于已有的计算机程序从而进行实际分析，这些程序通常提供了默认的标准选项，使用者会依赖它们，直到一些改动被认为是必需的。更重要的是，随着研究者逐渐熟悉因子分析的多种选项，他们会发现大多数的差别在很大程度上都只是表面的。实际上，有一些共同的线索贯穿于这些差别。更重要的是，研究者会发现，对相同的数据运用不同的方法和标准，在大多数实际应用中其结果是一致的。简而言之，读者没有必要立刻学会和使用所有的选项。然而，重要的是使用者要知道进行因子分析时最重要的差别和选项，并且使用者从一开始就要明白，对大多数问题我们没有绝对唯一的(或者最好的)解决方案。

本书假定读者对因子分析的概念基础有基本的认识，那些内容在上一卷书中已经介绍过。我们也希望读者明白以下两种内在的不确定性之间的区别，即从观察到的协方差结构推论潜在的因果结构的不确定性[①]（逻辑问题），以及从估计的样本统计量推论总体参数的不确定性（统计问题）。尽管这两个问题在解决实际的因子分析问题时有内在的联系，但在头脑中要分清这两者之间在概念上的差别。在讨论统计方法与实际问题之前，我们有必要简单复述一下第一卷所介绍的内容。

对因子分析基础的回顾

因子分析假定我们观察到的变量是一些潜在的（假设的或者未被观察到的）因子的线性组合。这些因子中有一些被假定是两个或更多变量共同的因子，另一些被认为是各变量独有的因子。这些独有的因子被进一步假定为（至少在探索性因子分析中）互成正交关系。因此，这些独有的因子对变量之间的共变关系没有贡献。换言之，只有共同因子（被认为在数量上要远比观察到的变量少）对观察到的变量之间的共变关系有贡献。

因子分析中假定的线性系统是这样的，如果我们知道潜在的因子负载，就可以准确无误地识别出协方差结构。然而，要从观察到的协方差结构中弄清楚潜在的共同因子结构却比较困难。这些基本的不确定性跟统计估计没有关系，且必须在统计以外的假定的基础上解决，它们包括因子因果关系的假定和关于简洁的要求。

在这些假定和线性系统的特性下，我们可以研究所得到的协方差结构，准确识别出潜在的因子模式，只要潜在的因子模式相对简单，且满足简单因子结构的要求。例如，图 1.1 所示的两个常见的因子模型在第一卷书中也展示过，它可以用表 1.1 中下三角区内无误差的相关矩阵恢复得到（在第一卷书中，我们指出用最大似然因子分析法得到结果后，再用基于直接最小斜交法的斜交旋转，

① 也就是变量之间有相关关系并不意味着有因果关系。——译者注

就可以产生与图 1.1 所示相同的因子负载。我们同时也指出,任何计算机程序,无论它们依赖何种运算法则,都应该能很好地得出类似的模式[1])。

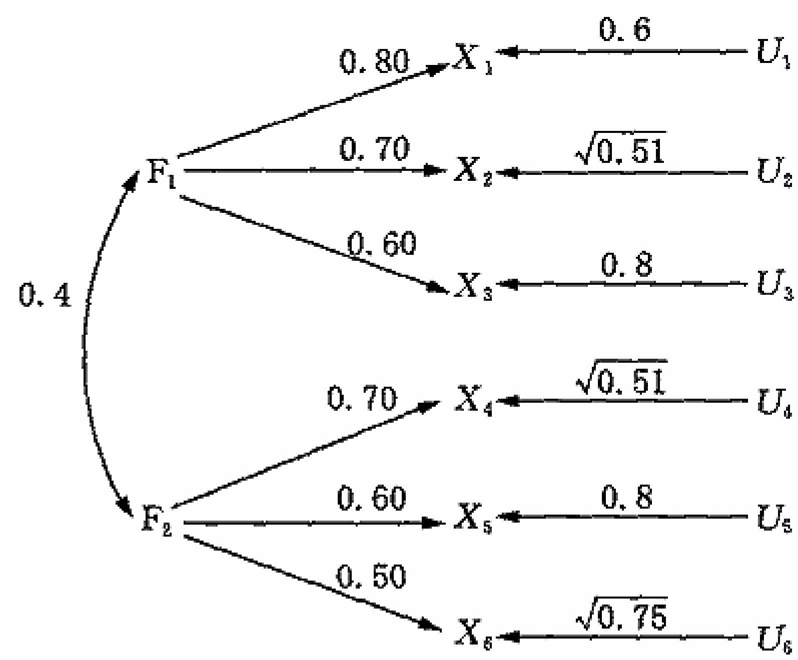


图 1.1　6 个变量的路径模型,2 个斜交因子模型的例子,其中可观察到的变量代表对以下事项的观点:

X_1 = 政府是否应该花费更多的钱在学校上;
X_2 = 政府是否应该花费更多的钱在降低失业率上;
X_3 = 政府是否应该控制大型企业;
X_4 = 政府是否应该通过校车接送学生来加速消除种族隔离;
X_5 = 政府应该保证少数族群能得到他们相应的工作配额;
X_6 = 政府应该扩展启蒙计划。

表 1.1　总体(下三角区)和模拟而得的 100 个样本(上三角区)之间的相关,对应于图 1.1 中的 2 个共同因子的模型[a]

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	—	0.6008	0.4984	0.1920	0.1959	0.3466
X_2	0.560	—	0.4749	0.2196	0.1912	0.2979
X_3	0.480	0.420	—	0.2079	0.2010	0.2445
X_4	0.224	0.196	0.168	—	0.4334	0.3197
X_5	0.192	0.168	0.144	0.420	—	0.4207
X_6	0.160	0.140	0.120	0.350	0.300	—

注:a. 由 University Paper 07-013 的表 8 和表 13 重新生成。

然而,实际上我们研究的协方差矩阵会被一系列随机或非随机的误差所影响,从而跟总体中的因子模式所暗示的协方差矩阵不同。为方便之后参考,我

们在表 1.1 的上三角区重新生成一个协方差矩阵，这个矩阵是由图 1.1 的因子模式所定义的理论总体中抽取 100 个样本而来(或者说是由表 1.1 的下三角区显示的协方差矩阵而得来)，注意到上三角区和下三角区相应数值的不同。此外，从相同总体中抽取的任何一个相关矩阵都会跟总体的相关矩阵有所不同，也会跟被抽取出的其他相关矩阵不同。因此，实际上不可能完全重现潜在的因子模式；我们只是尝试对潜在数值进行估计，让这些估计符合某些统计的和(或)实践的标准。

研究者通常用三个步骤来获得探索性因子分析的方案：(1)准备合适的协方差矩阵；(2)抽取初始的(正交的)因子；(3)旋转以获得最终的因子。最后，我们在相关的另一卷书中给出了如何通过多种软件包来实施上述步骤，并且强调要获得基本的因子分析信息是比较简单的。

本书涵盖的基本方案和方法

因子分析主要用于探索或者验证，取决于研究者的主要目的。在两种应用中，准备合适的协方差矩阵、抽取初始的因子、旋转以获得最终结果这三个步骤都会涉及。虽然要获得最终结果并不总要遵循这些步骤(尤其在检验特定的假设时)，但通过这些步骤来讨论因子分析的主要变型会比较方便。因此，本书的第一部分主要围绕这些步骤来展开。

在第一卷书中，我们已经指出在选择基本的录入数据时有一个重要的选项——是用普通的变量之间的协方差(或相关关系)，还是用各实体的相似描述。迄今为止，我们的讨论都集中在前者，此处也会这样。

在因子分析的最初步骤中，我们有共同因子模型，它作为我们的参照模型，也有主成分分析，它背后的原理跟“共同”因子分析不同。但它们都是广为运用的有效方法，都是探索变量之间“相互依赖关系”的工具。这两种方法的基本差别，在于主成分是观察到的变量的某种数学函数，而共同因子则无法用观察到的变量的组合表达出来。在因子分析的最初步骤中，另一种方法是映像因子分析法。映像分析跟共同因子分析的不同之处在于，共同因子分析把观察到的变

量看做是潜在无限的总体中的一个样本,而映像因子被定义为变量的线性组合。本书会更详细地解释这些方法的异同。此外,如果用的是共同因子模型,实际上也有很多方法来抽取最初的因子。在本书中会介绍到的抽取方法有:(1)最大似然抽取法(包括饶氏正则因子分析法),(2)最小二乘法(包括最小化残差法和带迭代公因子的主轴因子分析法),(3)Alpha 因子分析法。Alpha 因子分析法可以看做是共同因子模型的一个变型,也可以看做是另一种方法。

因子旋转这个步骤包括两个主要选项,即正交旋转和斜交旋转。斜交旋转又可以继续细分为对因子模式矩阵的负载进行直接简化,以及对参照轴的负载进行间接简化。在以上各选项中又有其他的变型。它们中的大多数将会在后面的章节中谈到。

接着,在下一章中我们将讨论应该抽取和保留多少个因子。本书在抽取方法的章节以外特意再安排这样的一章,目的在于介绍一些"经验法则",以方便实际应用者。

验证性因子分析这章的内容很基础。我们会给读者介绍一些一般性的用实证材料来检验因子分析模型的概念,并提供两个简单但很重要的验证性因子分析的例子。

然后,我们会讨论如何建立因子尺度以用做其他研究。我们把它放在验证性因子分析之后,是因为建立因子尺度在某些方法上的不足可以通过验证性因子分析来缓解。

在最后一章中,我们用一种问答的方式涵盖了许多问题。这些问题要么在正文中没有提及,要么非常重要,所以值得复述。在这一章我们也会对一些尚未形成共识的问题提供实践中的建议。

本书最后的名词解释不是为了给各个术语提供精确的技术定义,而是为了方便说明它们在本书中是在怎样的上下文情景中被使用。最后,本书的参考文献不是为了反映这个领域的历史发展,或对其创新给予足够的称赞。我们引用了我们对该领域所了解的有价值的文献,读者应该以同样的精神来使用这些文献。我们希望没有冒犯那些对因子分析的发展作出了贡献,却没有被我们引用的学者。

第 2 章 | 抽取初始因子的方法

在探索性因子分析中,抽取因子这个步骤的主要目的是决定共同因子的最小数量,这些共同因子能够产生可观察到的令人满意的变量间的相关关系。如果没有测量误差和抽样误差,且在数据中对因子因果关系的假定是合适的,那么在产生一个相关矩阵所需要的共同因子的最小数量和调整后的相关矩阵的秩之间就会有完全对应的关系(调整相关矩阵要求在其主对角线上插入公因子)。也就是说,在没有抽样误差,以及因子模型跟数据完全吻合的情况下,我们可以通过研究调整后的相关矩阵的秩得到公因子(实际值,而不是估计值),也可以得到共同因子的数量。然而,如果存在抽样误差,就不能依靠秩理论了。这时候的目标是在存在抽样误差的情况下,用一些准则来决定共同因子的数量。正如在《因子分析导论:它是什么以及如何运用》中讨论过的,决定共同因子的最小数量的最终准则,是这些被选定的共同因子能在多大程度上重现观察到的相关关系。因此,这个目标可以被重新表述为解决一个统计问题,即寻找一个决定我们何时停止抽取共同因子的准则。遵循标准的统计逻辑,它涉及到在什么情况下,我们认为重新生成的相关关系和观察到的相关关系之间的差别可被归结为抽样变异性。

我们从描述在多种抽取方法中一致的基本方案开始,其中包括假设用于重新生成观察到的相关关系所需的共同因子的最小数量。在没有任何信息的情况下,这意味着从一个共同因子的模型开始。这个“假设”通过运用某些标准来判定选取的模型与数据之间的差异是否微不足道。如果不是,我们将增加一个共同因子,运用同样的准则再对模型进行估计。这个过程将不断持续,直到它们之间的差别被认为可被归结为抽样误差(读者应该回顾《因子分析导论:它是

什么以及如何运用》的第 2 章，在那里我们假定没有抽样误差，其准则就是选取的模型与数据之间的差别为 0)。需要指出的是，实际运用的运算法则也许不能真正地进行这种连续的判断，但抽取初始的 k 个因子以解释我们观察到的共变关系的这个原则仍然是成立的。

虽然这个基本方案在原则上很直接，但它的应用却可以有无数种形式，因为有很多最大拟合(或最小差异)的标准。跟我们迄今描述的共同因子模型非常一致的两种主要方案是：(1)最大似然法(Lawley and Maxwell, 1971; Jöreskog, 1967; Jöreskog and Lawley, 1968)，它的变型包括正则因子分析法(Rao, 1955)，以及基于最大化残差偏相关矩阵的行列式方法(Browne, 1968)；(2)最小二乘法，它的变型包括迭代公因子的主轴因子分析(Thomson, 1934)和最小化残差法(Harman, 1976)。此外，还有另外三种主要的抽取因子方法：(1)Alpha因子分析法(Kaiser and Gaffrey, 1965)，(2)映像分析(Guttman, 1953; Harris, 1962)，(3)主成分分析(Hotelling, 1933)。我们将首先描述最后提到的这个方法。

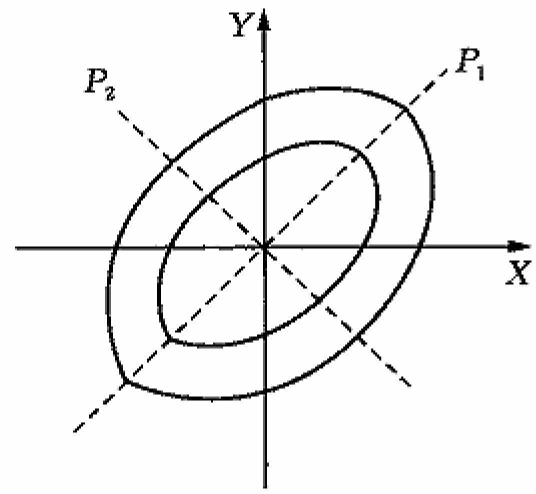
主成分，特征值和向量

我们之所以从讨论主成分分析开始，有两个原因。第一，它可以作为一个基本模型，我们可以用它和共同因子模型进行比较；第二，通过它很容易描述一些难以解释的概念，例如特征值和向量，以及它们在因子分析运算中的角色(我们并没有放弃依赖最简单的运算法则这个目标，但是对这些术语有一定程度的熟悉，对许多计算机程序的使用者来说是非常必要的。在此强烈建议读者看完以下的基本介绍)。

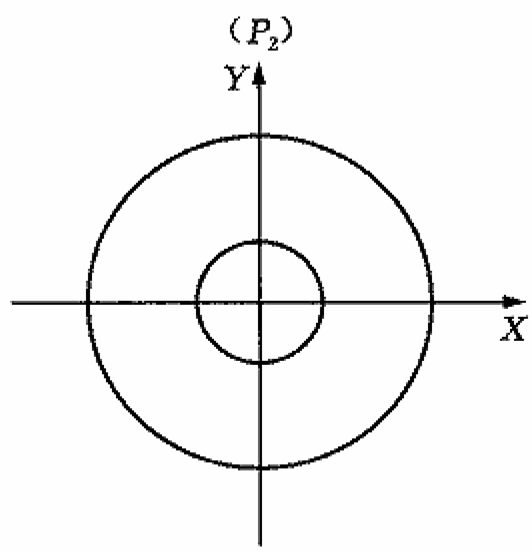
主成分分析是将我们观察到的一组变量转换为另一组变量的方法。为了展示它的特点和潜在的逻辑，最简单的方法是在双变量的情境下进行研究。假定有两个变量，X 和 Y，为了介绍上的方便，我们也假定它们的分布是二元正态分布。

图 2.1 用轮廓图描述了中等程度的正相关的二元正态分布关系。这些图显示，因为 X 和 Y 有正相关关系，数据点的集群表现为 X 的较大值与 Y 的较大值相关(反之亦然)，因此大多数个案都集中在第一和第三象限，而不是第二和第四象限。这些轮廓图形成椭圆形，它的两条轴用虚线表示。主轴(P_1)通过大多

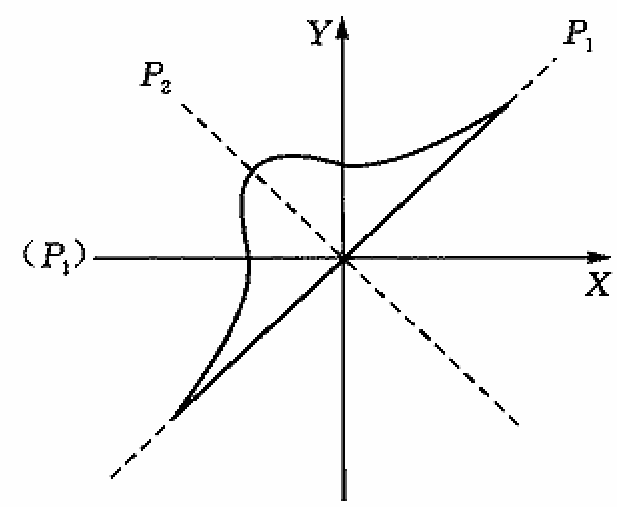
数数据点所在的那条线；次轴（P_2）通过最少数据点所在的那条线。



a. X 和 Y 存在某些关系



b. X 和 Y 没有关系



c. X 和 Y 完全相关

图 2.1　两变量分布的主轴的例子

假设我们现在的工作是只用一个维度或一条轴来表示每个个案的相对位置。对这个参照轴的合理选择是 P_1，因为从整体上来看，在某种意义上该轴跟数据点最接近。第一个主成分因而只是主轴上的个案的代表。举例来说，如果个案在 X 和 Y 上取值都为 1，在 P_1 上将表示为大于 1，而在 P_2 上则表示为小于 1。如果我们对每个个案都用 P_1 和 P_2 表示（用坐标系中的点），就不会有信息上的损失，从而我们可以对每一个案进行准确描述，而且不必考虑 X 和 Y 的关系。然而，当 X 和 Y 的关系变得更强时，我们可以说第一条轴（和第一个成分）在描述个案方面信息量更大。在极端情况下，如果 X 和 Y 是线性函数关系，那么第一个主成分就包含了描述任何一个个案的所有信息。如果 X 和 Y 之间是独立的，就不会有主轴，使用主成分分析就不能帮助简化问题。

尽管我们用椭圆和二元正态分布来表示主轴,但主轴的概念并不仅限于正态关系。一般来说,主轴用一条线来表示,各数据点到这条线的距离的平方和是最小的。与最小二乘原则的比较也许能帮助我们解释这一点。在寻找最小二乘回归线 ($\hat{Y}=a+bX$) 的过程中,我们把 Y 和 $\hat{Y}$ 的距离的平方最小化,也就是最小化 ($Y-\hat{Y}$),这里所说的距离由一条平行于 Y 轴且垂直于 X 轴的线来测量。在寻找主轴的时候,我们最小化数据点和轴之间的垂直距离(即点到主轴的垂直线,而不是点到 X 轴的垂直线)。这个区别我们用图 2.2 表示(Malinvaud, 1970,对最小二乘法和正交回归的讨论)。

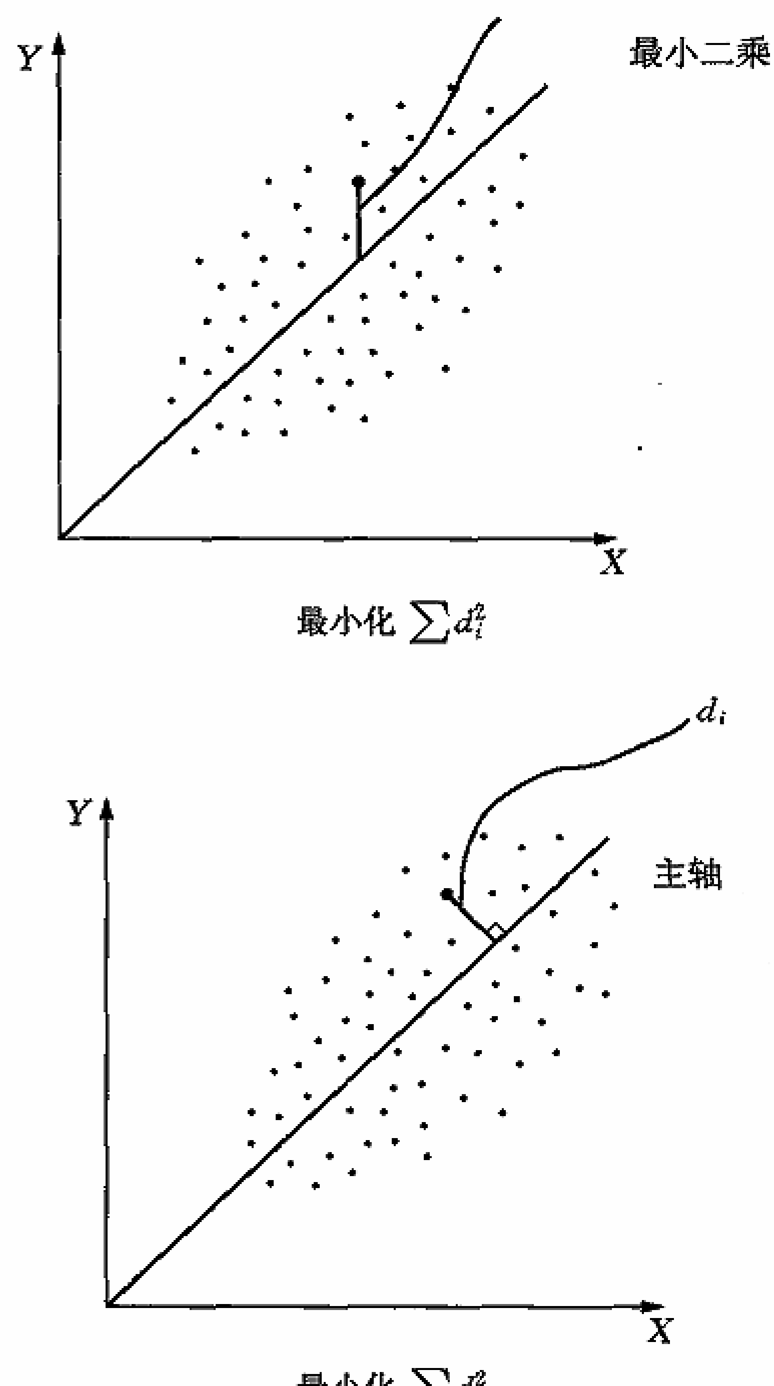


图 2.2 最小二乘回归和主轴的比较

一旦第一个成分被定义为包含最多的信息（它解释了数据中最大的方差），第二个成分也用相似的方式定义，此外还有一个条件是它要垂直于第一个成分。因此，在两变量的情况下，一旦确定了第一个成分，我们就自动得知了第二个成分。另外要记住的是，除非 Y 是 X 的线性函数，或者相反，否则就有两个主成分（我们需要用两条轴来完整地描述该联合分布）。

在获得主成分的过程中，我们不需要假定存在假设性的因子。新的轴是观察到的变量的数学函数。尽管主成分分析是用于获得表现上的简约（在只研究前面几个成分来说是如此），但它的目的不是解释变量之间的相关关系，而是尽可能解释数据中的方差。另一方面，因子分析性的分解只要求一个因子（在此两变量的情况下），且主要的目标是解释变量之间的相关关系。总而言之，前者（主成分分析）以解释方差为目的，后者（因子分析性的分解）以解释协方差为目的。

当有两个以上的变量时，界定主成分的基础是一样的。举例来说，在三元正态分布中，三维的轮廓图会与一个有些许扁平的足球相似（椭球）；第一条主轴穿过一个顶端到另一个顶端，因而成为最长的线；第二条轴穿过第二长的距离，同时垂直于第一条轴（这是由于扁平而变宽的部分）；第三条轴将是最短的，它会穿过由于扁平而变窄的部分。

为完成这样的层级分解或转换的数学方法被称为特性方程或特征方程。解这个方程将得到跟一个矩阵有关的特征值和特征向量。这个特性方程有以下的形式（用矩阵符号表示）：

$$RV = \lambda V \qquad [2.1]$$

其中 R 是我们要获得的矩阵，V 是我们要确定的特征向量，λ 是一个特征值。解决方法最终要靠以下形式的更简单的行列式方程：

$$\mathrm{Det}(R - I\lambda) = 0\text{，它将转变为一个双变量矩阵} \qquad [2.2]$$

$$\mathrm{Det}\begin{pmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{pmatrix} = 0 \qquad [2.3]$$

它也可以被写做

$$(1-\lambda)(1-\lambda)-r_{12}(r_{12})=0 \quad [2.4]$$

（根据行列式的定义）

$$=\lambda^2-2\lambda+(1-r_{12}^2)=0 \quad [2.5]$$

（将之展开并按照标准形式合并）

如果你记得怎么解方程 $ax^2+bx+c=0$，就可以得到特征值。在任何情况下，两变量相关矩阵的特征值为：

$$\lambda_1=1+r_{12}\text{，且} \quad [2.6]$$

$$\lambda_2=1-r_{12}\text{。} \quad [2.7]$$

注意，如果两个变量之间是完全相关的，则其中一个特征值为 2，另一个为 0；如果两个变量之间完全不相关，则两个特征值都为 1。

此外，注意特征值的和等于变量数，即 $\lambda_1+\lambda_2=(1+r_{12})+(1-r_{12})=2$，特征值的积等于相关矩阵的确定因素，即 $(\lambda_1)(\lambda_2)=(1-r_{12}^2)$。这些属性对各种大小的相关矩阵都成立。然而，最重要的是，最大的特征值代表第一条主轴所解释的方差有多少，第二大的特征值代表第二条轴所解释的方差有多少，以此类推。因为(使用相关矩阵时)特征值的和等于纳入分析的变量的数量，将第一个特征值除以 m(变量数)，我们也可以得到给定的轴或成分所解释的方差的百分比：

$$\text{一个成分所解释的百分比}=(\text{相应的特征值})/m \quad [2.8]$$

为得到相关的特征向量，我们需要强制性地增加一个约束条件，即它们的长度为 1。因而，主成分的负载可通过把特征向量乘以相应特征值的平方根得到，这就正确地反映了相应数据所解释方差的多少。

《因子分析导论：它是什么以及如何运用》中表 2 和表 3 提供了一个两变量(一个共同因子)的模型，它是最简单的主成分分析的例子。特征值(按降序排列)是 1.48 和 0.52，它们分别等于 $(1+r_{12})$ 和 $(1-r_{12})$。相关的特征向量分别

为($\sqrt{1/2}$，$\sqrt{1/2}$)和($\sqrt{1/2}$，$-\sqrt{1/2}$)。“因子负载”则表示为：

$$\begin{pmatrix}\sqrt{1/2} & \sqrt{1/2}\\ \sqrt{1/2} & -\sqrt{1/2}\end{pmatrix}\begin{pmatrix}1/\sqrt{1.48} & 0\\ 0 & 1/\sqrt{0.52}\end{pmatrix}=\begin{pmatrix}0.86 & 0.51\\ 0.86 & -0.51\end{pmatrix}$$

如果在计算机程序中使用主成分分析，可以得到最后的矩阵的因子负载。注意，$\lambda_1=(0.86)^2+(0.86)^2$，$\lambda_2=(0.51)^2+(-0.51)^2$。

为便于后面对成分分析和共同因子分析进行比较，作者把成分分析用于表1.1中所示的6个变量的相关矩阵。为了强调成分分析在没有抽样导致的波动的情况下的特性，我们使用无误差的数据。表2.1展示了成分分析的结果。在此要指出3点：(1)一般来说会得到6个成分(最后4个成分不重要，因而此处不报告)；(2)前2个成分比前2个共同因子解释的方差更多(61.6%对41%)；(3)前2个成分不能完全解释所观察到的相关关系，但前2个因子可以(例如，$(b_{11}b_{21})+(b_{12}b_{22})=(0.747)(0.706)+(-0.395)(-0.409)=0.6890$，大大高于潜在的相关系数0.56)。

主成分分析跟因子分析相似，它们都是简化数据的方法。研究者可以根据特征值的大小决定是否只用前2个成分。在此重复一下，尽管两个共同因子的模型可以解释所观察到的相关关系，但实际上并不必要。如果这两种方法被看做是探索变量之间互相依赖关系的方法，那么它们之间会有另一个相似之处。请注意，如果变量之间没有相关关系，就不会有主成分，因为任何一个成分都跟另外的一样，不分主次；各成分都只能解释一个单位的方差。随着变量之间相关关系的增大，被前几个成分所解释的部分也会增加。

区分这两种方法的一种方式是说因子分析用一个假设性的因果模型代表了协方差结构，而成分分析通过对观察到的数据进行线性组合总结了数据。在两者间如何选择取决于研究者总的目标。通过强加一个假设的模型，我们可以用较少数量的因子来解释相关关系。把观察到的数据的线性组合用数学方式表达出来，并不需要强加某些人认为有问题的因果模型，但它也没有揭示出任何潜在的因果结构，如果这样的结构确实存在。

表 2.1　表 1.1 下三角区的相关矩阵的前两个主成分

变　量	主　成　分		
	h^{2a}	F^1	F^2
X_1	0.749	−0.395	0.713
X_2	0.706	−0.405	0.666
X_3	0.651	−0.417	0.597
X_4	0.595	0.579	0.623
X_5	0.548	0.529	0.581
X_6	0.488	0.526	0.514
特征值	2.372	1.323	和 = 3.695
被解释的方差的百分比	39.5	22.1	
被解释的方差的累积百分比	39.5	61.6	

注:a. 这些并不是严格意义上的公因子估计,因为主成分分析并不假定共同因子的存在。

因此,成分分析的目的跟因子分析有很大的不同。然而,我们还是希望重复一下为什么我们要花费那么多篇幅在这上面。首先,主成分分析法经常被认为是因子分析的一种变型。其次,(后面会讲述的)主轴因子分析运用相似的运算方法(特征方程),如果有成分分析的知识,介绍因子方法会更容易。最后,最重要的是,成分分析所产生的一个统计量,将作为实际运用最广泛的、解决因子数量问题的方法(它指的是"特征值大于 1"的标准,我们将在后面谈到这个问题)。

共同因子模型的变型

从历史上说,以前对因子分析的大多数解释性处理都是用主轴因子分析的步骤来确定共同因子模型,它将主成分分析的分解方法运用在调整后的相关矩阵上,它的(单位为 1)对角线上的元素被公因子的估计值所代替。

常用的公因子的估计值是各变量跟组内其他变量的多重相关系数的平方,或者在相关矩阵的一行中绝对值最大的相关系数。把这些公因子代入相关矩阵的对角线后,因子就用主成分分析的方法被抽取出来了。也就是说,跟成分

分析一样，我们通过把相同的特征值方程应用在调整后的相关矩阵上，就得到了因子分析的结果(这也是主轴因子分析这一名称的由来)。在这种情况下，所要解的方程式为：

$$\mathrm{Det}(R_1 - \lambda I) = 0, \qquad [2.9]$$

在这里 R_1 是相关矩阵，它的对角线上是公因子的估计值。尽管这种方法被广泛使用，但它逐渐被后面介绍的最小二乘法所取代。

最小二乘法

用最小二乘法进行共同因子分析背后的原则是在抽取一定数量的因子后，最小化残差相关系数，并且评估用模型重现出来的相关关系跟观察到的相关关系之间的拟合程度(通过研究两者之间差别的平方)。由于人们总是可以通过假定跟变量数相等的因子数来重现观察到的相关关系，并且因为总是可以通过增加假定的因子数来提高拟合程度，因此在最小二乘法中，我们只需要假设 k 个因子(k 比变量数要少)就能重现观察到的相关关系。

获得解决方案的实际过程大致是这样的。首先，假设用 k 个因子就可以解释观察到的相关(这一步实际上并没有提出任何独特的问题，因为我们可以从一个共同因子的假设开始，逐渐增加假设的因子数，直到我们得到一个满意的答案)。第二，得到一些公因子的初始估计值(正如前面所讲到的，我们将使用一个变量跟其他变量之间的多重相关的平方)。第三，(根据最小二乘法原则)获得或者抽取最能重现观察到的相关矩阵的 k 个因子。在这一步，要解的方程跟前面的方程 2.9 完全一样。第四，为了获得能最好地重现观察到的相关矩阵或协方差矩阵的因子模式，我们根据上一步得到的因子模式重新估计公因子(《因子分析导论：它是什么以及如何运用》第 2 章的方程 20 就是估计公因子的公式)。最后，这个过程被不断重复，直到我们不能进行任何改善。这就是带迭代估计公因子的主轴因子分析法的名称由来。

最小化残差或 Minres(Harman, 1976)也是基于相同原则的一种迭代方法，

但它使用了一种在某种程度上有所不同却更有效率的运算法则。在大样本的情况下它有一种近似于卡方检验的技术。这种检验方法独立于抽取因子的方法,Harman 提出这个近似的检验方法可以应用于其他抽取方法上,且可以作为检验因子分析是否完全的方法(Harman, 1975: 184; McDonard, 1975)。尽管这个检验在大样本的情况下是合适的,但讽刺的是,当样本量较大时,细微的差别也会在统计上显著。因此,Harman 建议不要只依赖这种正式检验,而应把通过这种检验得到的因子数量仅仅作为上限的标志,(最好在研究了旋转结果之后)尽量只保留那些有实际意义且理论上可以被解释的因子。

对表 1.1 中的相关矩阵使用带迭代的主轴因子分析法的结果展示在表2.2 中。

表 2.2 带迭代估计公因子的主轴因子分析法:政治态度的例子

变 量	F_1	F_2	h^2
X_1	0.731	−0.320	0.637
X_2	0.642	−0.282	0.492
X_3	0.550	−0.241	0.360
X_4	0.513	0.473	0.487
X_5	0.441	0.409	0.362
X_6	0.367	0.340	0.251
特征值	1.842	0.746	
被解释的百分比	30.7	12.4	

基于最大似然法的解决方案

最大似然法解决方案的总目标和最小二乘法一样:寻找能够对观察到的相关关系进行最好拟合的因子组合。对这个原则的非正式表述如下:我们假定观察到的数据构成了总体中的一个样本,一个有 k 个共同因子的模型跟这个总体完全吻合,且在总体中这些变量(包括因子)是多元正态分布的。然而,被假定为未知的是参数的确切构成,即每个变量的确切负载。于是目标就成为(在给定的假设下)寻找潜在的总体参数,使之产生我们观察到的相关矩阵的可能性最大。一个基于相同原则但又稍微不同的标准是寻找假设性的因子构成,使 k

个共同因子和观察到的变量的正则相关系数最大。第三个标准最后也指向相同的准则，即寻找使残差相关矩阵的行列式最大化的因子结构。所有这些标准在现实应用中都很复杂，基于相同原则的各种方法在通过迭代进行转换过程中的效率也有很大差别。目前 Jöreskog(1967)的解决方案被公认为是最好的。

下面我们将证明，这个步骤在原则上和其他特征方程方法并没有大的不同。它的基本运算法则可以通过之前讨论过的行列式方程的形式表示：

$$\det(R_2 - \lambda I) = 0, \qquad [2.10]$$

此处 R_2 由以下方程得到

$$R_2 = U^{-1}(R - U^2)U^{-1} \qquad [2.11]$$

$$= U^{-1}R_1U^{-1} \qquad [2.12]$$

此处 U^2 是在各个步骤中独有方差的估计值。方程 2.10 和方程 2.4 的不同之处在于，它用调整后的矩阵 R_2 代替 R，并在每次迭代中重新调整；它和最小二乘法的不同之处在于，R_2 在每个步骤中都被调整，且赋予有较小的独有方差的相关关系以更大的权重。注意方程 2.11 中 $(R-U^2)$ 这部分跟方程 2.9 中的 R_1 一样，唯一不同的是方程 2.11 或 2.12 中的加权因子。在最大似然法中，独有方差被看做是“半”误差方差，因此这个方法给有更大公因子（或有更小独有方差）的变量以更大的权重，这符合统计估计的一般性效率原则，即给不那么稳定的估计以更小的权重。

我们之前提到过，最优的方法应该可以在模型被界定好、且数据没有误差的情况下准确地重现总体值。但有一些程序不一定能得到这样的结果，这取决于特定程序的收敛效率；然而原则上说，好的程序应该可以做到。对表 1.1 上三角区的样本相关关系运用最大似然法，且假设有两个因子，我们将得到表 2.3 的结果。

正如我们所预料的，显著性检验显示它进行了足够的拟合。下面列出的计算卡方的公式只是为了说明这个值取决于样本量，而自由度独立于样本量。卡方统计值由以下的方程决定：

表 2.3 对表 1.1 上三角区的数据运用最大似然法和两个共同因子的方法

变量	未旋转的		公因子	用直接最小斜交标准旋转后	
	F_1	F_2		F_1	F_2
X_1	0.747	−0.300	0.648	0.817	−0.027
X_2	0.701	−0.266	0.562	0.754	−0.009
X_3	0.599	−0.176	0.389	0.602	0.046
X_4	0.428	0.362	0.314	0.027	0.547
X_5	0.505	0.605	0.621	−0.113	0.833
X_6	0.534	0.248	0.367	0.202	0.468
平方和[a]	2.132	0.749		1.652	1.215
有 4 个自由度的卡方 = 0.825					

注:a. 平方和与未旋转的结果中的特征值相等,将这个值除以 m 得到被该因子解释了的方差的百分比。在斜交旋转的结果中,它们所代表的只是可能被称为各因子的"直接"贡献。总贡献(包括由于因子之间相关而导致的贡献)仍然等于未旋转的结果中的特征值之和。

$$U_k = N\{\ln|C| - \ln|R| + tr(RC^{-1}) - n\} \qquad [2.13]$$

$\ln$ = 自然对数,且 tr = 矩阵的迹;

N = 样本量;

n = 变量数;

R = 协方差矩阵;

$C = FF' + U^2$,此处 F = 因子负载且 U^2 为独有方差。

(实际上,检验最小二乘法用的是相同的公式,唯一不同的是对 F 和 U 的不同估计。)重要的是对一个固定的相关矩阵来说,U_k 的值随着 N 按比例增大。相应的自由度由以下公式得到

$$df_k = 1/2[(n-k)^2 - (n+k)]。 \qquad [2.14]$$

在这里 k 是假设的因子的数量,n 是变量数。注意 df_k 不会受到样本量 N 的影响。

这个方法最大的优点在于,它提供了对大样本的显著性检验。如果卡方检验显示观察到的数据明显偏离有 k 个共同因子的模型,我们就可以决定 $k+1$ 个共同因子的模型是否合适。在探索性分析中,我们通常会从有一个共同因子的

假设开始，直到显著性检验表明我们得到的模型没有显著地偏离观察到的数据。尽管这些连续的检验并不互相独立，而且使用它们的时候也不必有太多顾虑。(Lawley & Maxwell, 1971)。

然而，在实际应用中，当样本量较大时，如果我们仅仅依赖于显著性检验，就会遇到一个问题，即我们得到的模型中的共同因子数量比我们预期的更多。此外，因子模型只是对现实的近似表现，所以模型和数据之间如果有拟合上的细微问题，就有可能产生一个新的显著的因子。在第 4 章中我们将会再次讨论关于如何决定因子数量的一些问题。

Alpha 因子分析法

无论是最小二乘法还是最大似然法，它们都假定我们考虑的变量构成了总体，而且涉及到的抽样只是对个体的抽样。然而，在 Alpha 因子分析法中，包含在因子分析中的变量被看做是变量总体的一个样本，同时假定这些变量在一个给定的由个体组成的总体中被观察到。因此，在 Alpha 因子分析中，重点是强调心理测量的推论，而不是通常的统计推论。

Kaiser 和 Caffrey 认为这个方法基于一个原则，即我们抽取出来的共同因子与相应总体中存在的共同因子有最大的相关性，由此而决定因子负载(1965:5)。

对这种方法的另一种理解是把独有的因子看做是心理测量抽样中产生的误差。由此，对公因子的估计被看做是在测量背景下的"信度"。作为第一步，这种方法首先产生一个修正了"衰减"后的相关矩阵：

$$R_3 = H^{-1}(R - U^2)H^{-1} \qquad [2.15]$$

此处 U^2 和 H^2 分别是独有成分和公因子的对角矩阵(H^{-1}是一个对角矩阵，它包含公因子的平方根的倒数)。然后与这个"修正了"的矩阵相关的行列式可以用通常的方程来解：

$$\det(R_3 - \lambda I) = 0 \qquad [2.16]$$

我们对方程 2.16 和方程 2.10,以及方程 2.15 和方程 2.11 做一些解释,发现它们之间的相似和不同之处是有启发性的。最大似然法用独有方差来调整矩阵,而 Alpha 因子分析法用公因子来调整矩阵。换言之,前者给有更大公因子的变量以更大的权重,而后者则相反。和之前一样,实际的解决方案是复杂的,因为我们要从初始的公因子开始,对它的值进行迭代直到得到最后的结果。

在 Alpha 因子分析法中,要保留多少个因子的标准是(与该因子)相关的特征值应该大于 1。这个标准等同于另一个标准,即与之相关的概化系数 α 在变量的总体中要大于 0(这种方法因而被称为 Alpha 因子分析法)。因此,这通常就不会有显著性检验,因为它假定我们正在考虑的就是所有个体构成的总体。

对表 1.1 的上三角区的样本相关矩阵运用 Alpha 因子分析法得出的结果参见表 2.4,该表同时也给出了我们后面要讲的映像因子分析法的结果。

表 2.4 对表 1.1 报告的无误差的相关关系进行 Alpha 和映像因子分析法得出的因子负载[a]

变量	未旋转的因子矩阵					
	Alpha			映像		
	F_1	F_2	公因子	F_1	F_2	公因子
X_1	0.669	0.437	0.638	0.575	0.133	0.348
X_2	0.586	0.384	0.490	0.538	0.139	0.309
X_3	0.502	0.329	0.361	0.477	0.131	0.245
X_4	0.585	−0.382	0.489	0.372	−0.270	0.211
X_5	0.502	−0.329	0.360	0.335	−0.263	0.182
X_6	0.419	−0.274	0.251	0.287	−0.239	0.140

注:a. 把这些值与 Kim-Mueller, University Paper 07-013 中的表 10 比较,在那个表中观察到的相关关系在此被完全重新生成。而且我们注意到,用 Alpha 因子分析法估计的公因子跟真实的公因子非常接近,而映像因子分析法对公因子的估计相对较差。

映像分析

映像分析区分了一个变量的共同部分和独有部分。一个变量的共同部分

被定义为由一组变量中除这个变量之外的其他所有变量的线性组合所预测的部分，它被称为该变量的映像。变量的独有部分是这个变量中不能被其他变量预测的部分，它被称为反映像。在进行这种分解时，我们假定处理的是变量的总体和个体的总体，同时没有对它们中的任何一个进行抽样。

映像分析也假定这个变量的总体是无限大的。为了便于比较，我们回到图1.1所界定的两个共同因子的模型。在那个模型中被界定的6个变量构成了某种意义上的总体。但是在映像分析中，这6个变量应该被视为一个无限大的变量总体中的样本，它跟两个共同因子所覆盖的心理测量维度有关。

然而，如果研究潜在变量总体中的所有变量，我们可以发现，一个变量的映像的平方就等于共同因子分析中所定义的变量的公因子，一个变量的反映像的平方就等于独有方差(在这里我们假定所处理的是标准化后的变量)。换言之，一个变量与该变量所属的变量总体中其余变量之间多重相关系数的平方跟该变量的公因子相等。

一个变量样本中的映像和反映像分别被称为偏映像和偏反映像。尽管偏映像只是近似于总映像，但它完全是被观察到的变量所决定的。从这个意义上讲，它和共同因子分析有很大的不同，在共同因子分析中变量的共同部分被定义为假设性因子的某种线性组合，且从来不是观察到的变量的确切函数。

给定一个变量的样本和它们之间的相关关系，映像分析能建构出一个偏映像协方差矩阵，它可以表示为

$$R_4 = (R - S^2)R^{-1}(R - S^2) \qquad [2.17]$$

在此 R 是相关矩阵，S^2 是对角线矩阵，它的元素是各变量中不能被其他变量解释的方差——或者说是反映像方差。在方程2.17中涉及到的过程是：(1)把R的对角线替换为各变量与其他变量的多重相关的平方；(2)重新调整对角线以外的元素，由此得到一个Gramian矩阵。然后把特征方程应用到以下的矩阵：

$$\det(R_4 - \lambda I) = 0 \qquad [2.18]$$

然而,要保留的因子数量并不是由分析方程 2.18 的特征值得到的,而是用另一个矩阵($S^{-1}RS^{-1}$)替代公式中的 R_4 后,根据特征值大于 1 得到。通常来说,由此得到的要保留的因子数量较多——接近于被分析的变量数的一半。Kaiser 建议在进行合适的旋转后,放弃那些不显著的和无法解释的因子。对样本相关矩阵进行映像因子分析后得到的一些统计值显示在表 2.4 中,该表同时也给出了另外一些结果。

第 3 章 | 旋转的方法

因子分析的第一步，通常是决定要充分解释观察到的相关关系所需的最少的因子数量，在这个过程中也决定各变量的公因子。因子分析的下一步是通过旋转寻找更简单和更容易解释的因子，同时保持因子数量和各变量的公因子不变。

上一章讨论的所有方法都可以生成初始因子，它们是垂直的关系，并且按照从最重要到最不重要的顺序排列。因子分析结果的这两个特点并不是数据结构中内在就有的；它们是对数据的强制限定，以使最后的结果唯一且在某种意义上可被限定。进行这些强制限定，其结果是：(1)无论潜在的真实模型怎样，变量的因子复杂性都可能大于 1，也就是说，变量会在超过一个因子上有相当程度的因子负载；(2)除了第一个因子，其他因子是有两极的，也就是说，在一个因子上某些变量有正的负载，而另一些变量有负的负载(如果读者看不懂这些描述的意义，可参看第一卷书的第 2 章，并考察一些因子模式)。

目前有三种方法可以解决(因子)旋转的问题。第一种方法是像我们在第一卷书第 2 章讲过的那样，通过作图研究变量的模式，然后旋转数轴或者定义新的数轴，使得新的数轴能最好地满足我们的标准，即结构简单又有意义。当存在明显的变量丛时，它们之间互相分离，如果使各轴都穿过一个变量丛，就可以得到简单的结构。但当模式并不那么清晰或者有很多因子需要考虑时，对初学者来说这样的图像旋转并不实用。

第二种方法是依靠一些分析性的旋转方法，至少在选择了是否达到简单化的判定标准后，它们不依赖于主观判断。这个方法有两种不同的亚类型——一种是正交旋转，另一种是斜交旋转。在各亚类型中又有很多不同的变型，但在本章中我们只介绍几种众所周知并广为应用的方法。

第三种旋转的方法是在实际的旋转前定义一个目标矩阵或构成。这种旋转的目标是寻找最接近给定目标矩阵的因子模式。因为对目标矩阵的设定要求一定程度的知识和对因子结构性质的假设,因此这个策略跟验证性因子分析相近。

图像旋转,简单结构和参照轴

当(变量)丛不明显,或者有两个以上的因子需要考虑时,我们就很难使用图像旋转的方法。我们谈及这个方法仅仅是为了在介绍分析性的旋转方法时有一个起点。读者可以参考 Mulaik(1972)的著作,书中对图像旋转进行了很详细的介绍。

所有旋转方法的显性目标都是为了得到可能的最简单的因子结构。但遗憾的是,简单性这个概念本身并不那么直接,所以它没有给我们提供一个正式的和无异议的标准。Thurstone(1947)对界定简单结构进行了最雄心勃勃的尝试,但现在我们一致认为他的标准并不一定都经得起分析性的考验。因为对 Thurstone 的标准的理解需要有超平面和子平面的知识。对那些理解向量空间的读者,我们只展示 Mulaik(1972:220)对这些标准的精彩介绍。在 Mulaik 的描述中,r 指的是共同因子的数量,V 是有一个参照轴的参照结构矩阵。

(1) 参照结构矩阵 V 在每一行至少有一个 0。这是在本章开头简单结构的定义所暗示的对简单结构的基本假定。

(2) 在参照结构矩阵 V 的每一个 k 列,应该有一组数量至少为 r 的线性无关的可观察到的变量,它们与第 k 条参照轴的变量的相关系数为 0(这可以在 V 的第 k 列找到)。为了超定相应的参照轴,这个标准是必要的。

(3) 对 V 的每两个列,其中一列中应该有几个 0,且另一列中相应的位置应不为 0。这个要求保证了参照轴的唯一性,以及它们对应的共同因子空间的 $r-1$ 维度的子空间的唯一性。

(4) 当得到 4 个或 4 个以上的共同因子时,V 的每两个列应该相应有一定比例的 0。这个要求保证每个参照轴都只跟一些观察到的变量相关,从而保证

把观察到的变量分到不同的丛中。

(5) V 的任何两列都只能有少数相应的位置不能为 0。这个标准进一步保证了变量的简单性。

这些标准从根本上都基于两个在某种程度上不同的考虑:(1)界定简单的因子结构标准的需要;(2)说明在各种情况下简单结构如何被清晰界定的需要。Thurstone 的标准对初学者来说难以理解,主要是因为关于第二个考虑的文献技术性很强且很复杂。然而,就我们的目的来说,第一个考虑是基础,第二个考虑代表的是技术上的要求,我们把它交给专家们去解决。

尽管很难说明什么构成了"简单结构"的最小要求,但对给定的 r 个因子和 n 个变量,要说明什么构成了可能的最简单的结构则相对容易。如果所有变量的因子复杂性都是 1——即各变量只在一个共同因子中有不为 0 的负载,因子结构就是最简单的。如果有 2 个或 2 个以上的共同因子,就意味着在最简单模式的矩阵中:(1)每一行都只有一个不为 0 的要素;(2)每一行都有一些 0;(3)对任何两个列,不为 0 的要素不会重叠。

在实际数据中我们很难遇到这样的简单结构。因此问题就变成了如何"界定"因子结构,使之"最接近"最简单的结构。在这个问题上,专家们对在"不完美"模式中界定"简单"结构,以及为了得到这个简单结构的计算方法上有不同的看法。如前所述,Thurstone 的标准界定了在什么实证情形下简单结构可以被清晰定义。其中一种实证情况,是至少有 3 个变量清楚地在各个因子上有负载。但简单结构的定义并不依赖于这种实证要求,且在数据分析中实际确定简单结构时最好撇开这个标准。事实确实如此,因为在探索性因子分析中,研究者必须满足于手头上既有的变量,且在尝试给予因子意义前,就被迫把构成最简单结构的因子概念化。

在历史上,最早是通过参照轴来界定简单结构的。尽管对它的了解并不是绝对必须的(因为不依赖于参照轴的斜交旋转方法的发展),但我们还是会简单地加以介绍,因为因子分析的许多使用者需要依赖计算机程序,而这些程序是基于参照轴来提供斜交旋转的方法。

第一卷书中提到,初始的因子负载不过是在两条轴上对变量进行投影——

也就是说,负载是通过对每个点往两条初始的正交轴上垂直画线,再读取其在两条轴上的交点而得到。要注意到,在这种正交方法中如果所有变量都在轴上,就可能出现简单结构。也要注意到,在正交的情况下,简单结构意味着一组点在另一条轴或因子上的负载为0(或投影为0)。如果丛之间的角度不是正交(即不为90°),投影为0就不可能。如果出现这么一个斜交的角度,新的步骤就是建立另一条垂直于超平面(在这个双因子的模型中,它只是一条线)的参照轴,它穿过被认为是主要因子轴的点丛(参看图3.1)。

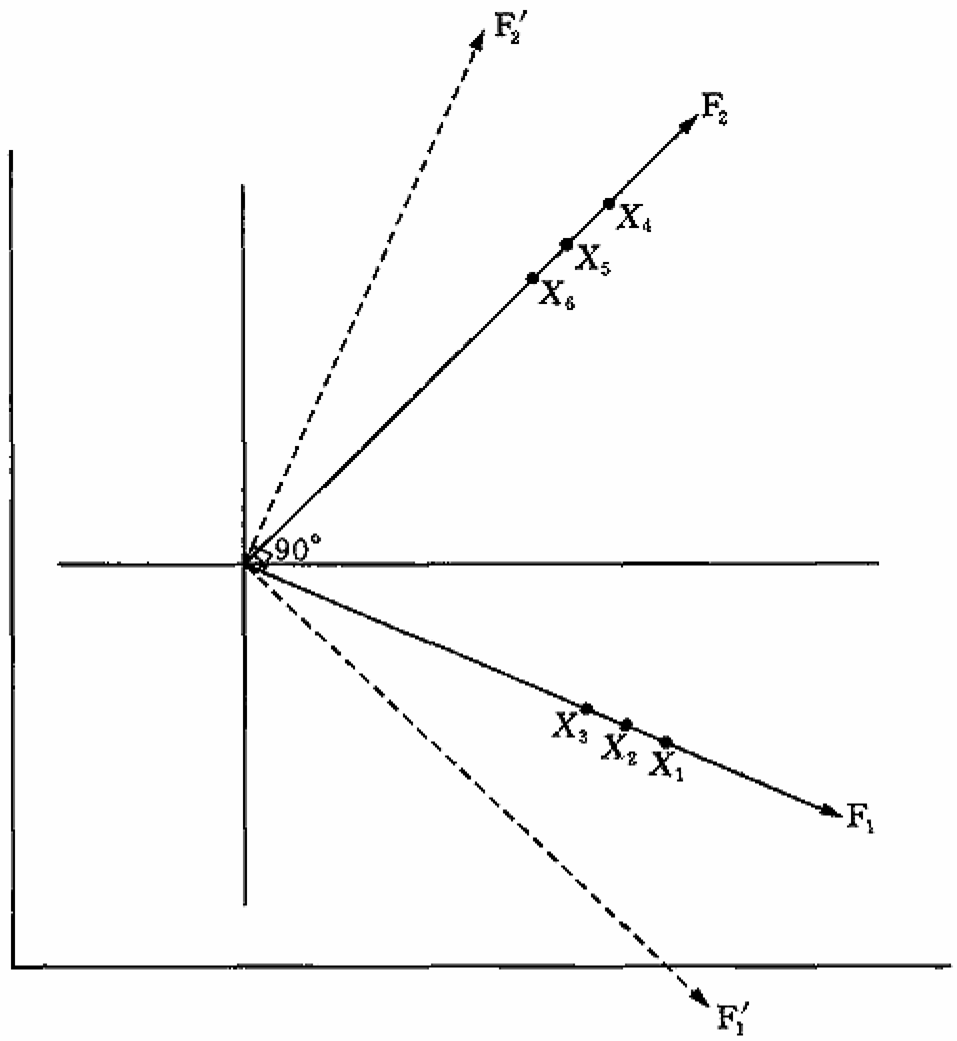


图3.1 F_1 和 F_2 是主要的斜交因子,F_1'和 F_2'是相应的参照轴。X_1, X_2 和 X_3 在 F_2'上的投影为0,且 X_4, X_5 和 X_6 在 F_1'上的投影为0

因此,研究一个变量丛都在一条主轴上的情况,和研究一个变量丛在参照轴上投影为0的情况是一样的。(如果是从无误差的相关矩阵来研究)在我们设计出来的两个共同因子的模型中,这种情况就是一组变量 X_1,X_2 和 X_3 在参照轴 R_2 上的投影或者负载为0,而另一组变量在 R_1 上的负载为0。在这个例子中,很难解释清楚为什么我们要依赖于参照轴,而不是直接画线穿过变量丛。

我们只是说明当存在 2 个以上的因子，且变量并不像人为数据那样明显呈丛时，确定参照轴的方法，并把这些参照轴在下一步或迭代中看做垂直的轴，从而能够帮助我们找到一个拟合得更好的主轴。在这里重要的是，记住旋转的基本目标始终是寻找一个因子模式矩阵，使之最接近于前面提到的最简单的理想结构。

正交旋转的方法：四次方最大法，最大方差法和均等变化法

由于我们假定读者在实际分析数据时要依赖于现有的因子分析的计算机程序，因此我们只描述每种方法背后的基本原则。在前一章中，我们介绍了在给定 k 个共同因子和 n 个变量情况下的最简单的结构。在这里总结一下这种矩阵的分析性属性是很有必要的。

因为一个变量只在一个因子上有负载，因此对变量的因子解释是最简单的。但那样的特点并不足以用数字表示它的简单程度。其中一种对简单程度的数学测量是每行(或每个变量)的因子负载的平方的变化程度(我们考虑负载的平方只是为了避免处理负载的正负号问题)。因为方差被界定为与均值的离差的平方的平均值，如果其中一个元素的负载的平方等于公因子，且行中的所有其他元素都为 0，(对固定数量的因子和固定的公因子来说)这时方差将是最大的。简而言之，一个变量的因子负载的平方的最大方差等于该变量因子复杂性的最简程度。因而，将因子简单程度的概念数量化是合理的：

$$\text{变量 i 的因子复杂性} = \frac{1}{r}\sum_{j=1}^{r}(b_{ij}^2 - \bar{b}_{ij}^2)^2, \qquad [3.1]$$

在此 r 是模式矩阵中列的数量，b_{ij}是变量 i 在因子 j 上的因子负载，$\bar{b}_{ij}^2$是行内因子负载的平方的平均值。公式 3.1 可以写成以下的形式：

$$q_i = \frac{\sum_{j=1}^{r}(b_{ij}^4) - (\sum_{j=1}^{r} b_{ij}^2)^2}{r^2} \qquad [3.2]$$

一旦得到了初始的因子结果，r 和各变量的公因子就固定了。因此，在正交方法

中减号后面的部分仍固定,是因为

$$\sum_{j=1}^{r} b_{ij}^2 = h_i^2$$

由此,简单程度的总测量就可以通过计算所有变量 q_i 的总和而得到:

$$q = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} \frac{\sum_{j=1}^{r}(b_{ij}^4) - (\sum_{j=1}^{r} b_{ij}^2)^2}{r^2} \qquad [3.3]$$

运用四次方最大标准会得到因子负载最小化 q 的旋转轴。然而,最小化 q 等于最大化以下项,

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{r} b_{ij}^4, \qquad [3.4]$$

因为公式 3.3 中的其他项都是常数。因此,它被称为四次方最大法。

实际上,运用这个标准可能导致强调变量解释的简单性,但损害了因子解释的简单性。尤其是在涉及较少共同因子的时候,一个变量的解释也更简单,而如果相对较少的变量在因子上有很高的负载,且其他变量在其上没有负载,因子的解释就更简单。总的来说,四次方最大原则倾向于某个一般性因子在一些变量上有中等或较少负载的情况下产生最终的结果。

最大方差旋转法用一种稍微不同的标准来简化因子矩阵的各列。它最大化各因子负载的平方的方差,而不是最大化变量负载的平方的方差。要最大化的量——因子 j 的简单性指标为:

$$v_j = \frac{n\sum_{i=1}^{n} b_{ij}^4 - (\sum_{i=1}^{n} b_{ij}^2)^2}{n^2} \qquad [3.5]$$

注意现在是计算 n 个变量的总和,且跟公式 3.2 中的相应项不同,减号后面的 $(\sum_{j=1}^{n} b_{ij}^2)$ 不固定。对其简单性的总测量为

$$V = \sum_{j=1}^{r} v_j = \frac{\sum_{j=1}^{r} n\sum_{i=1}^{n} b_{ij}^4 - (\sum_{i=1}^{n} b_{ij}^2)^2}{n^2}, \qquad [3.6]$$

它被称为行最大方差标准。在迭代过程中，通常我们用正态化后的因子负载以减少初始负载对最后的结果影响过大的问题。在公式 3.6 中我们用 b_{ij}^2/h_i^2 代替 b_{ij}^2，用 b_{ij}^4/h_i^4 代替 b_{ij}^4 就可以得到这个标准。

在表 3.1 中我们给出了用四次方最大法和最大方差法（正态化后）对相同的数据进行旋转的结果。在这里我们要指出，尽管四次方最大法在分析性上比最大方差法更简单，但最大方差法对因子进行了更清晰的区分。总的来说，Kaiser 的实验(1958)表明，对不同子集的变量进行因子分析时，用最大方差旋转法得到的因子模式比用四次方最大法得到的更稳定。

表 3.1　对表 2.3 展示的相同的模式矩阵运用最大方差和四次方最大旋转法[a]

变　量	最大方差旋转法		四次方最大旋转法	
	F_1	F_2	F_1	F_2
X_1	0.787	0.167	0.793	0.133
X_2	0.730	0.170	0.736	0.143
X_3	0.595	0.187	0.602	0.166
X_4	0.154	0.539	0.173	0.533
X_5	0.083	0.783	0.111	0.780
X_6	0.306	0.503	0.324	0.492

注：a. 在这个例子中，四次方最大旋转法得到的第一个因子是“一般性”因子的可能性很低。

考虑到四次方最大标准专注于简化因子矩阵的行，而最大方差标准专注于简化因子矩阵的列，因此逻辑上我们也可以考虑同时使用这两个标准，并给它们合适的权重。一般的标准可以表示如下：

$$\alpha Q + \beta V = \text{最大}, \qquad [3.7]$$

在此 Q 由公式 3.4 得到，V 由公式 3.6 得到（但为了处理上的便利要乘以 n，因为乘以一个常数不会影响最大化的过程），α 和 β 是权重。这可以写成

$$\sum_{j=1}^{r}\left(\sum_{i=1}^{n} b_{ij}^2\right)^2 - \gamma\left(\sum_{i=2}^{n} b_{ij}^2\right)^2/n = \text{最大}, \qquad [3.8]$$

在此 $\gamma = \beta/(\alpha+\beta)$。

如果 $\gamma = 0$，它就成为四次方最大标准；如果 $\gamma = 1$，它就变成最大方差标

准。当 $\gamma = r/2$，它被称为均等变化法；当 $\gamma = 0.5$，它被称为二分四次方最大法。

斜交旋转的方法

斜交旋转比正交旋转应用得更广泛，因为它不对因子强加限制，即不要求它们不相关。它相对于正交旋转的优势在于，在进行斜交旋转后，如果得到的因子是正交的，我们就可以确定其正交特性并不是人为设定旋转方法的结果。然而，因为斜交法引入了因子之间的相关，因此对因子分析进行解释时另一种复杂性又出现了。尤其是我们可能要假定高阶的因子因果关系来解释因子之间的相关。此外，斜交旋转法有两种不同的方法——一种利用参照轴，另一种利用主模式矩阵。因为我们在前面的章节里讨论了获得简单结构的总原则，所以在此对它们的描述将尽可能简单。

基于参照轴的方法

这里涉及的所有方法都基于这样一个事实：存在可界定的变量丛，它们代表不同的维度，且如果这些丛被主因子正确地界定出来，各丛变量只会在一条参照轴有投影，而在其他参照轴上投影为0。因此，跟四次方最大法的标准类似，我们可以定义四次方最小标准：

$$N = \sum_{i=1}^{n} \sum_{j<k=1}^{r} a_{ij}^2 a_{ik}^2, \qquad [3.9]$$

在此 a_{ij} 和 a_{ik} 是在第 j 条和第 k 条参照轴上的投影。如果所有变量都是单因子的，这个值会是0。但我们在斜交旋转中要得到的是能够最小化N的因子负载。在正交旋转中，这个标准等同于四次方最大法。

对应于正交旋转法中最大方差法对四次方最大标准的修改，这里有最小协方差或二分四次方最小标准。在这里要最小化的值是在参照轴上的投影的平方的协方差：

$$C=\sum_{j<k=1}^{r}\left(n\sum_{i=1}^{n}a_{ij}^{2}a_{ik}^{2}-\sum_{i=1}^{n}a_{ij}^{2}\sum_{i=1}^{n}a_{ik}^{2}\right) \quad [3.10]$$

同样地，如果我们用 a_{ij}^2/h_i^2 代替 a_{ij}^2，我们就能得到一个基于正态化值的结果。如果应用到相同的数据上，最小协方差标准倾向于产生更少的斜交因子，而四次方最小标准倾向于产生更多的斜交因子。

由于两种标准有相反的倾向，因此结合它们就很自然。最一般化的方法为：

$B=\alpha N+\beta C/n=$ 最小，在此 α 和 β 是要加的权重，N 和 C 由前面的公式得到。 [3.11]

将公式 3.11 乘以 n，合并同类项，且设 $\gamma=\beta/(\alpha+\beta)$，我们得到一般的最小斜交标准：

$$B=\sum_{j<k=1}^{r}\left(n\sum_{i=1}^{n}a_{ij}^{2}a_{ik}^{2}-\gamma\sum_{i=1}^{n}a_{ij}^{2}\sum_{i=1}^{n}a_{ik}^{2}\right)。 \quad [3.12]$$

当 γ 的取值不同时，这个一般化的标准有变化：

当 $\gamma=0$（斜交最明显）时，为四次方最小标准；

当 $\gamma=0.5$（斜交不那么明显）时，为二分四次方最小标准；

当 $\gamma=1$（斜交最不明显）时，为最小协方差标准。

这里再提醒一次，通常我们用 a_{ij}^2/h_i^2 代替 a_{ij}^2，使用正态化后的最小斜交标准。

另一个与发展最小斜交法过程中界定的原则紧密相连的标准，是二分正态最小标准，但它基于完全不同的算法。它尝试对公式 3.12 中的 γ 进行更客观的选择，修正四次方最小法“过于斜交”和协方差最小法“过于正交”的偏误。如果数据特别简单或特别复杂，相对于选择 γ 为 1/2 的二分四次方最小法，二分正态最小法被认为更理想。

基于因子模式直接进行最小斜交法的方案

近年来，Jennrich 和 Sampson（1966）提出了一个标准，它基于简化主因子（而不是参照轴）上的负载，并成功地开发出计算机程序。要最小化的值跟公式

3.12 所定义的一致,唯一不同的是用主因子负载(在模式矩阵中的负载)代替参照轴上的负载。这个标准是:

$$D = \sum_{j<k=1}^{r}\left[\sum_{i=1}^{n} b_{ij}^2 b_{ik}^2 - d\left(\sum_{i=1}^{n} b_{ij}^2 \sum_{i=1}^{n} b_{ik}^2\right)/n\right] \qquad [3.13]①$$

在此 b_{ij} 是模式矩阵中的因子负载,D 和 B 之间的细微差别是在 D 中有除以 n。跟间接的最小斜交法一样,研究者可以通过选择公式 3.13 中的 d 来修改最终结果的斜交程度。

一般来说,如果 d 值越大,产生的结果斜交程度就越大;d 值越小(为负),则产生的结果正交程度越大。如果因子模式是单因子的(可能的最简单的模式),$d = 0$ 就能界定正确的模式。

在这里有必要提醒一下。公式虽然相似,但我们并不清楚在直接的最小斜交法中 d 的选择与间接的最小斜交法中 γ 的选择之间的对应关系。为了更清楚地了解界定不同的 d 的影响,读者可以参考 Harman 的书(Harman, 1975)。

其他斜交旋转法

还有很多其他的斜交旋转法,在此我们简单介绍一些更广为人知的方法。

最大斜交标准(Saunder, 1953)尝试通过最大化小的和大的负载的数量,牺牲中等负载,从而简化结构。作为一种客观的标准,它的方法是寻找将双倍的因子负载的峰度最大化的结果(这个翻倍通过将各负载计算两次而达到,即按负载的原符号计算一次,按负载相反的符号再计算一次)。这个标准和正交旋转法中的四次方最大法一样,但它在不限制数轴正交的情况下会得到与四次方最小法(四次方最大法在斜交旋转里的对应方法)不同的结果。

同时还要提一下另外两种旋转方法。目前被广泛应用的 Harris 和 Kaiser 的正斜交旋转法(Harris & Kaiser, 1964),及另一种可能的选择——最大平面

---

① 原书中为 $D = \sum_{j=k=1}^{r}\left[\sum_{i=1}^{n} b_{ij}^2 b_{ik}^2 - d\left(\sum_{i=1}^{n} b_{ij}^2 \sum_{i=1}^{n} b_{ik}^2\right)/n\right]$,怀疑有误。——译者注

法(Cattel & Muerle, 1960; Eber, 1966),这种方法基于一个跟前面提到的所有方法都不太相同的拟合标准。

旋转至目标矩阵

除了运用旋转方法得到有分析性定义的简单结构外,有时候更理想的是旋转因子以拟合特定的、研究者头脑中已有的或者假设存在的简单结构。

第一种可能是,研究者给定了在各变量上的确切负载,根据最小二乘标准,通过限制或者不限制正交来进行旋转,使给定的矩阵和旋转后最后得到的因子矩阵之间的差别[①]最小。这种旋转通常用来研究一种因子结构和另一种已知的或在其他地方展现过的因子结构之间的一致性。

第二种方法被称为最优斜交旋转法,是一种通过利用正交旋转结果的某种函数作为目标矩阵来得到斜交旋转的方法(Hendrickson & White, 1964)。最优斜交旋转法背后的原理是,正交旋转的结果通常跟斜交旋转的结果相似,通过将小负载减少到接近 0,我们可以得到一个相当好的简单结构的目标矩阵。然后通过给这个目标矩阵寻找拟合得最好的斜交因子,我们就可以得到理想的斜交旋转结果。另外还有不同的获得目标结构矩阵和目标模式矩阵的运算方法,但在这里我们不进行介绍。

第三种也是更常用的方法,是利用一个没有前述的目标矩阵那么明确的目标矩阵。我们不是通过最小二乘法拟合一个所有值都设定好的目标矩阵,而只是设定 0 和 1。通常我们不知道具体的值,但知道哪些负载会更高,哪些负载会更低,因此这是最现实的目标矩阵。表 3.2 给出了这样的目标矩阵的例子。

这样的目标矩阵可以通过修改使它更一般化:有一些值可以设定为 0,有一些则设定为其他值,其余的让它们自由变化。这些会在验证性因子分析的章节中更详细地介绍。

---

① 原文为拟合最小,怀疑有误。——译者注

表 3.2 有 0 和 1 的目标矩阵

变量	因子 1	因子 2
X_1	1	0
X_2	1	0
X_3	1	0
X_4	0	1
X_5	0	1
X_6	0	1

第 4 章 | 再论因子数量的问题

尽管我们之前研究过初始因子分析的几种方法,其目的是寻找符合数据的最小数量的因子,但是有几个原因促使我们重新研究这个问题。第一,在讨论初始因子分析方法时,我们实际上假定了可以无异议地解决因子数量的问题,因此也没有讨论跟这个问题相关的技术点。第二,部分初始结果并不是回答这个问题的理想工具,因此需要重新研究旋转后的结果。第三,我们不得不解决由于因子分析模型和数据之间不完美的拟合所造成的问题。第四,大多数现有的计算机程序会要求使用者提供关于因子数量问题的信息,我们希望为读者应对这种情况作好准备。

解决因子数量的问题通常有几个原则。其中一些是替代性的选项,另一些则是可以互补的。最重要的原则是:(1)与最大似然法和最小二乘法相关的显著性检验;(2)各种特征值准则;(3)实质重要性的准则;(4)碎石检验;(5)可解释性和恒定性的准则。

检验显著性

纯粹从统计角度来说,如果方法所要求的假定都成立,那么与最大似然法相关的大样本卡方检验是最理想的方案。应用这个方法的结果显示,如果样本量很大,有很多变量,最后保留的因子数量往往远比研究者预期的因子数量大很多。尽管这确实算不上这个方法的缺陷,但它迫使研究者在找到统计上重要的因子后,依赖另一个准则,即实质重要性原则。

蒙特卡洛实验显示,最大似然准则在已知的没有实质上不重要小因子的总

体模型上运用最合适。也就是说,它是解决抽样变异性有效的方法,但如果模型设定了小的偏误,它就不是最好的方法。只要样本量足够大,任何这些偏误都会被看做是"重要的"维度,并且不能被抽样变异性所解释。因此,这意味着在进行了合适的旋转后,以实质意义为依据来忽略不那么重要的因子也许更好。

尽管我们把这种方法描述为一个保险的步骤——它检验单个共同因子的模型是不是足够,如果数据"显著地"偏离单因子模型,我们就检验双因子模型是否足够,依次类推——但这个步骤用在大量变量上则很费劲。因此,我们可以把其中一种决定共同因子数量的快速方法(下面将会介绍)和最大似然法的检验结合起来。在一开始"猜测"了显著的因子后,如果数据明显地偏离假设的模型,我们可以增加因子的数量,或者如果初始的模型被认为足够而被接受,我们可以减少因子的数量(以保证我们获得符合数据的最少数量的因子)。从统计的角度来讲,最小二乘法没有最大似然法那么有效率,我们对检验显著性的方法也持有相同的看法。

通过特征值进行设定

最常用的解决因子数量问题的准则是在分解(未调整的)相关矩阵时保留特征值大于 1 的因子。这个简单的准则似乎很有效,它通常会产生与研究者的预期相一致的结果,它被应用到人为创造出来的总体模型的样本中也很有效。

在总体的相关矩阵中,这个准则总能设定共同因子数量的下限。也就是说,用于解释相关矩阵的共同因子的数量总是等于或者大于这个准则所界定的数量。然而,当我们考虑样本相关矩阵时,这个不相等的关系不一定能保持。尽管 Kaiser 对它的有效性给出了很多解释,但接受这一准则仍然是基于探索性和实用上的考虑。在研究了其他更"复杂的"准则后,Kaiser 仍然坚持使用这个准则(Kaiser, 1974)。

另一个相关的特征值准则是在分解简化后的相关矩阵(把多重相关系数的

平方放在对角线上)时保留特征值大于0的向量。这个方法背后的原理是,在总体相关矩阵中,它对用于解释数据的共同因子的数量规定了一个更为严格的下限。但用在样本相关矩阵上时情况可能不同,应用到实证数据上通常会产生多于依据其他考虑所能接受的因子数量。

在估计出公因子并将其插入到对角线上后,我们就可以运用这个特征值准则了。然而,如果有一些特征值为负(这种情况经常会出现),抽取所有特征值大于0的因子就没有什么意义。尽管正的和负的特征值的和与所有公因子的和(或者共同因子所解释的方差)是一样的,但我们无法在方差的背景下有意义地解释负值,而且他们的存在使得正的特征值之和“膨胀了”,因为它们的和大于公因子的和。Harman(1975:141)建议,在特征值的累积和大于所估计的公因子之和时就应该停止抽取共同因子。

实质重要性的准则

考虑到“显著性”检验专注于抽样变异性,特征值准则专注于矩阵的一些抽象属性,我们用的第三种准则是直接考虑因子所应有的最低贡献,否则这个因子就会被认为在实质上是不重要的。当初始因子分析是基于对未调整的相关矩阵进行分解时,这个准则很容易理解;要设定的是要保留的最后一个因子所解释的总方差的比例(即变量的数量)(回顾我们讨论过的抽取因子的各种方法,初始因子都是按照其重要性的顺序排列的)。我们可以随意设定被认为实质上很重要的准则的水平。其中一些选择是1%,5%,或者10%。但是,注意“特征值大于1”的准则等同于将能解释的方差比例的最小值设定为百分比(100/n)。

另一方面,在我们讨论过的各种方法中,除了主成分分析法,在对调整过的相关矩阵进行因子分析时,要界定的部分是最后一个特征值相对于特征值之和(要进行因子分析的矩阵的对角线之和)的百分比。这个方法最大的劣势在于它运用了主观准则,最明显的优势是不熟悉矩阵特征值分解属性的研究者可以基于相对重要性的测量来做解释,这种测量“看起来”也更容易解释。

碎石检验

这是 Cattell (1965)所推荐的检验。这个方法指示我们去研究特征值的图,在特征值(或特征根)开始变平并形成一条几乎是平行斜率的线时停止抽取因子。Cattell 称超过该点的平滑斜线为"因子垃圾或碎石"(碎石在此是一个地理学术语,指在石块坡的下面所找到的碎石)。图 4.1 给出了它的用法。基于这些结果,研究者总结出我们最多只能抽取 5 个因子。

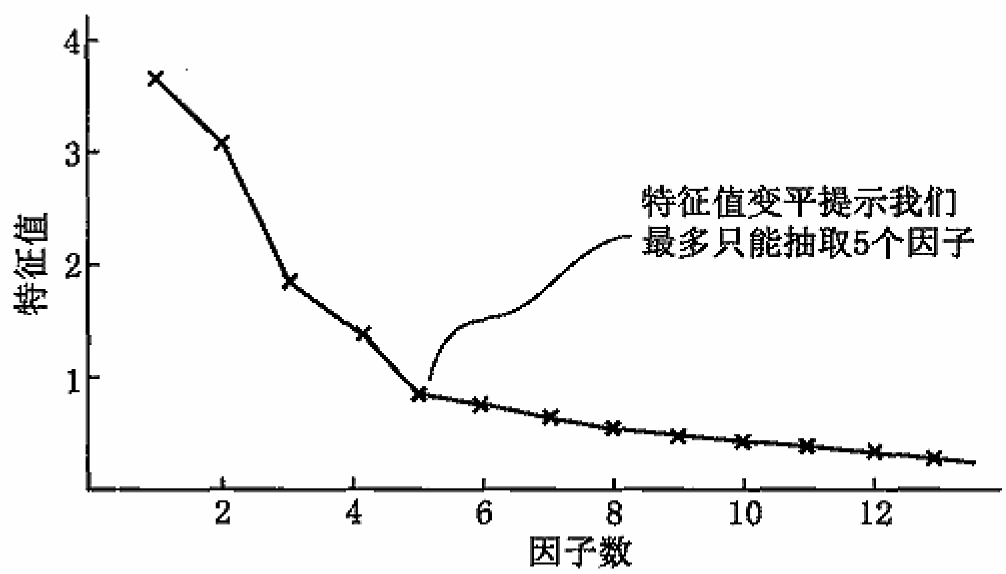


图 4.1　碎石检验的演示

蒙特卡洛的一些研究显示,当研究的兴趣只是寻找主要的共同因子,而次要的因子也存在时,这种方法往往比其他方法更优越(Tucker, Koopman & Linn, 1969; Linn, 1968)。有一些研究者,例如 Kaiser(1970),认为这种"观察特征根的"准则往往太主观,因为我们经常在特征根图中发现不止一个主要的断点,但却没有一个清晰的原则来让我们作出选择。

可解释性和恒定性的原则

为了避免得到一些不可信的结果,一般的基本原则是尝试结合多种规则,但只接受被多种独立的准则所支持的结论,而把其他的结果看做是暂时的假设

(Harris, 1967)。考虑到这种方法的复杂性和内在的不确定性,最后决定依据研究者所在领域中的现有学术水平来判断结果的合理性。这个准则难以捉摸,但或幸运或不幸的是,我们所有人都不得不接受这种情况,目的在于使我们的研究结果可以和其他科学家沟通。

第 5 章 | 验证性因子分析简介

迄今为止我们都专注于探索性因子分析，并且也强调为了运用这种技术，我们需要进行很多种假设，其中最重要的是因子因果关系和简洁性的假设。在这种分析中我们所做的都是给数据强加某一特定模型，寻找一个和数据最吻合的结果，我们可以合理地质疑，运用这种方法是否或者在多大程度上为因子分析性模型自身提供了实证支持。正如之前所述，无论结果如何，我们都不可能通过对协方差结构的观察来证明某种因果结构的存在。然而，我们可以估计因子模型在多大程度上被经验性地证实。

因子分析性模型被经验证实的程度

与探索性因子分析相比，我们在验证性因子分析中介绍了更多关于因子结构的特有的假设，所以如果一些因子因果关系不存在，这些特有的假设被一个给定的协方差矩阵所支持的可能性就更小。从这个意义上说，大多数验证性因子分析可以提供自我检验的信息。如果一个给定的因子假设被数据支持，我们也会有更充分的信心认为，对这个数据而言该因子分析性模型比较合适。（当然，在多大程度上被经验所证实会因不同的分析而异。）此外，即使是一个完全探索性的分析，它也能给我们提供不同程度的自我检验的信息。因此，在介绍验证性因子分析之前，重要的是讨论被一般经验证实的概念，以及一些用于判断我们的数据是否适合进行因子分析的方法。

一个例子

把因子分析性模型应用在双变量的相关关系上不会产生研究者已知以外

的信息。这主要是因为一个单个共同因子的模型总是和双变量相关关系相吻合。因此,因子分析从不在这种情况下运用,这不是因为因子分析性模型与数据不吻合,而是由于模型的经验证实度(以下简称为资讯性)为 0,同时也因为(尽管更微不足道)没有唯一的结果。考虑图 1.1 展示模型中的前两个变量。给定任何相关关系,我们都可以随意选择其中一个因子负载在 −1 和 1 之间(0 除外),且还能找到另一个因子负载使得它吻合观察到的相关关系。简单来讲,总是有和数据吻合的因子分析结果。

当因子分析被应用在三变量的相关矩阵上时,情况有些许改变。如果我们发现一个单个共同因子的模型与数据吻合,这时经验证实度不等于 0,因为一些随机的相关矩阵不能和一个单个共同因子的模型吻合。而且,如果一个三变量的相关矩阵和一个单个共同因子的模型吻合,三个相关系数之间的关系就必须满足两个条件:(1)所有相关系数都是正的,或者负的系数的数量必须是偶数;(2)任何一个系数的大小要等于或大于剩下的两个系数的积:

$$|r_{ij}| \geqslant |r_{ik} r_{jk}|。 \qquad [5.1]$$

导出公式 5.1 所界定的条件是有意义的。考虑《因子分析导论:它是什么以及如何运用》中的图 18(或者图 1 的上半部分),回忆

$$\begin{aligned} r_{12} &= b_1 b_2 \quad & h_1 &= b_1^2 \\ r_{13} &= b_1 b_3 \text{ 和 } & h_2 &= b_2^2 \\ r_{23} &= b_2 b_3 \quad & h_3 &= b_3^2 \end{aligned}$$

接下来,让我们把两个相关系数相乘:

$$r_{12} r_{13} = b_1 b_2 b_1 b_3 = b_1^2 b_2 b_3 = h_1^2 r_{23}。 \qquad [5.2]$$

第一个等号后面的项是用因子负载来表示其相关关系,第二个等号后面的项经过了重新调整,最后的项是给因子负载代入相应的公因子和相关系数。对公式 5.2 做些许调整,且因为公因子不能大于 1,我们得到公式 5.1 所界定的条件:

$$h^2 = r_{12} r_{13} / r_{23} \leqslant 1,$$

它暗示 $|r_{23}| \geqslant |r_{12}r_{13}|$。一般来说,所有三变量的单个共同因子的模型都必须满足条件 $|r_{ij}| \geqslant |r_{ik}r_{jk}|$。由于不是所有随机生成的三变量的相关矩阵都满足上述条件,所以我们的数据符合一个单个共同因子模型是有意义的,但并不是非常有意义,因为有很多随机矩阵也和单个共同因子的模型吻合。换言之,有很多三个相关关系的随机矩阵符合公式 5.1 界定的条件。

在《因子分析导论:它是什么以及如何运用》这本书的第 2 章我们讲过,一个基于单个共同因子模型的四变量的相关矩阵要符合三个额外的条件,即

$$r_{13}r_{24} = r_{14}r_{23} \qquad [5.3]$$

$$r_{12}r_{34} = r_{14}r_{23}$$

$$r_{13}r_{24} = r_{12}r_{34}$$

这个原则很容易导出和记住,举例来说,因为 $r_{13}r_{24} = b_1b_3b_2b_4 = (b_1b_4)(b_2b_3) = r_{14}r_{23}$(相同的步骤在导出公式 5.1 中的不等关系中也使用过)。所以,一般来说,变量的数量越大,为了符合一个特定的因子模型,相关矩阵需要满足的条件就越多。因此,如果一个单个共同因子的模型和一个四个变量的矩阵相吻合,它就给研究者提供了一定程度的经验证实,说明因子分析模型并不完全是随意的。

于是,因子分析结果就可以提供一些对模型自身是否合适的经验证实,因为只有当相关矩阵符合一些限制时,一个给定的因子模型才能和数据吻合。此外,变量数对假设的因子数的比率越大,因子分析模型的经验证实程度就越高。因为它也暗示为了满足因子分析性模型,在相关矩阵中会存在更多的结构限制。

我们说过,在运用因子分析时要对数据强加各种假设。因此,我们可以仅仅根据这些假设的随意性或不合适而否定因子分析性模型。然而,在经验证实度很高的时候,我们要小心进行判断,因为我们要考虑到数据本身可能也存在结构限制(即缺乏随机性)。从另一种不同的角度看,我们可以说,在不同的应用中因子分析的资讯性不同;一些因子分析的结果比另一些结果资讯性更高。再有一种看待它的角度是因子分析能够提供自我检验的信息;一个给定的结果

要满足的经验限制越多，我们对因子分析模型应用在数据上的合适性的信心就越大。因此，根据结果，即使是探索性因子分析也能提供模型是否合适和是否简洁的经验证实信息。

模型所暗示的经验限制的数量

由前面的讨论，我们明白，重要的是要知道一个给定的因子模型所暗示的经验限制的数量（为了拟合一个特定的模型，一个相关矩阵要满足的条件的数量）。幸运的是，这个数量等于最大似然法的显著性检验的自由度。（对这个数量进行详细研究是有意义的，由此我们可以对因子假设和它暗示的自由度有更清晰的了解，这对理解接下来要讨论的验证性因子分析非常重要。）

然而，有很多种方法可以用来确定一个特定的因子模型所暗示的限制的数量。其中一个方法是用秩理论，它指出，如果把正确的公因子放入相关矩阵（由一个有 r 个共同因子的模型产生），调整后的相关矩阵的秩（或者独立维度的数量）将会是 r，因此又暗示所有的超过 r 列和 r 行的子矩阵的行列式数量为 0。由此，我们可以对给定的因子数量和变量数量确定一个相关矩阵要满足的限制（Harman, 1976）。另一个方法是在显著性检验的背景下研究自由度。因为第一种方法得到的条件数量和运用显著性检验所确定的相同，因此我们将在第二种方法的更一般的背景下讨论。

为了举例，假设我们从一个经验性的相关矩阵开始。这个矩阵所包含的独立信息量为 $(1/2)n(n-1)$ ——上三角区的单元格的数量。给定这样的数据，因子分析通过允许 $n\times r$（其中 r 是共同因子的数量）的因子负载变化，使之最佳地重现观察到的相关矩阵，由此得到初始的结果。但在初始的因子分析中，我们要求这些 $n\times r$ 的因子负载满足一个条件，即最后得到的因子是正交的。这暗示强加了 $1/2r(r-1)$ 个条件。因此，为得到初始因子结果，我们有以下数量可自由变动的参数：

$$nr-(1/2)r(r-1) \qquad [5.4]$$

所以,在模型中一个经验性的相关矩阵要独立满足的条件数量是以下两者之差:

$$\text{所要求的经验限制的数量} = 1/2n(n-1)-[nr-1/2r(r-1)] \quad [5.5]$$
$$= 1/2[(n-r)^2-(n+r)],$$

这和前面展示的自由度一样(当我们用一个协方差矩阵代替一个相关矩阵时,矩阵包含的独立信息量是 1/2n(n+1),而不是 1/2n(n-1)。然而,最后的自由度仍然相同,因为要增加额外限制,使得因子分析模型可以在这个协方差矩阵上运用)。

表 5.1　与 n 个变量和 r 个因子相关的自由度[a]

变量数(n)	因子数				有正的自由度的因子的最大数量	独立系数的数量 $\frac{1}{2}n(n-1)$
	1	2	3	4		
3	0	−2	−3	—	无	3
4	2	−1	−3	−4	1	6
5	5	1	−2	−4	2	10
6	9	4	0	−3	2	15
7	14	8	3	−1	3	21
8	20	13	7	2	4	28
9	27	19	12	6	5	36
10	35	26	18	11	5	45
11	44	34	25	17	6	55
12	54	43	33	24	7	66
20	170	151	133	116	14	190
40	740	701	663	626	31	780

注:a. 一般化的公式为 $D=\frac{(n-r)^2-(n+r)}{2}$ = 数据要满足的限制的数量。

为了熟悉公式 5.5 所暗示的内容,我们在表 5.1 中给出了多种因子数量和变量数量组合的实际值。有几点要指出:第一,一般来说,随着变量数对因子数的比率的增加,要满足的经验限制的数量也在增加。第二,如果经验限制的数量为负值,因子分析结果就不能对模型提供任何经验证实。因此,一般来说只考虑对数据有一些经验限制的因子模型是有意义的。举例来说,对一个有 4 个

变量的矩阵运用 2 个共同因子的模型，或者对一个有 6 个或 6 个以下数量变量的矩阵运用 3 个共同因子的模型，都是没有意义的。第三，因子数保持不变，随着变量数的增加，所要求的限制会迅速增加。因此，在因子分析中每增加一个变量，就可以对因子分析性结果增加很多经验性的内容。第四，如果我们把这个数字作为经验证实度的指数，它暗示真正重要的不是比率，而是变量数和假设的共同因子数之间的差。需要注意的是，要满足的限制的数量在以下组合中几乎相同：1 个因子和 7 个变量(14)；2 个因子和 8 个变量(13)；3 个因子和 9 个变量(12)，等等。但没有理由认为这个指数是经验证实度的直接测量。另一个可以考虑的选项是(a)要满足的限制数对(b)观察到的矩阵中独立的系数数量的比率。尽管我们没有把这些比率显示在表中(基数在最后一列)，但需要指出的是，这个测量所显示变量数对因子数的比率会比这个粗指数所显示的更重要。

在评价因子分析结果的经验证实度时，要考虑两个问题。第一个是这些要求即使在总体中完全得到满足，也可能在样本中无法完全满足。第二个是即使在总体中，因子分析性模型也未必完全和数据吻合。因此，在评价这个要求时要考虑些许不拟合的情况。但最让人沮丧的是在实际分析中没有办法把这两个问题分离开来。因此，实际上我们不能只用公式 5.5 作为对经验证实度的测量。给定观察到的数据和因子分析结果之间的一定程度的拟合，一个需要满足更多经验限制的结果的证实度就更高。但为了进行评价，我们要寻求一种方法来测量拟合度。

经验证实度或信度

和一些初始因子分析方法相关的显著性检验评价了假设的模型和观察到的数据之间的差异在多大程度上可以归结为抽样误差。显著性检验直接依赖于样本量，只要样本量够大，模型和观察到的数据之间的任何差异都会变得显著。其原因在于，如果因子模型和总体数据完全吻合，样本量越大，总体值和样本统计量之间的差别就越小。如果样本量非常大，那么离差就要非常小。

当研究者怀疑存在次要的因子，但不愿意或者无法确定它们的结构时，这

个统计原则就会出问题。在这种情况下,显著性检验不一定能反映模型充分与否——也就是说,即使被确定的因子模型很简洁,即它能解释大部分观察到的协方差,给数据结构带来了一些规则,检验也可能会显示该模型在统计上是不充分的。因此,我们需要一个描述因子模型是否充分的指数,它在概念上要和统计显著性不相关。

我们需要的统计量是对观察到的相关矩阵和重制的矩阵之间差别的测量。Harman 建议,一种方法是用残差均方——它把观察到的相关关系(根据因子分析的最后结果)和预测到的相关关系之间离差的平方加总——除以我们要考虑的单元格的数量:

$$\sum_{i \neq j}\sum (r_{ij} - \hat{r}_{ij})^2/[n(n-1)],$$

此处的加总是对所有不在对角线上的元素加总(Harman, 1976:176)。然而,这个测量没有一个便于参考的上限,所以我们很难解释它的相对大小。

另一个测量是 Tucker 和 Lewis 为最大似然因子分析法设计的信度系数(1973)。这个测量依据的是剔除最后一个因子的影响后矩阵内剩余的相关系数。因此,它最终依据的是观察到的相关关系和基于因子分析结果的相关关系之间的拟合程度。然而,他们的信度系数整合了另两项调整:它把总差别除以自由度,因此调整了因子分析结果之间的潜在差异,且它比较了调整后的离差以及与之可比较的假设因子不存在情况下的离差,因此使它成为削减调整后的误差比例的一种测量。因此,这个系数取值在 0 和 1 之间,前者意味着最差的拟合,后者意味着完全拟合。公式的简化形式是:

$$rho = \frac{M_0 - M_k}{M_0} \qquad [5.6]$$

在此 M_0 等于在没有因子作用下的预期的 χ^2 除以 $(1/2)n(n-k-1)$ ①,M_k 等于最后结果的 χ^2 除以 $(1/2)[(n-r)^2-(n+r)]$ (Sörbom & Jöreskog, 1976:4—5)。当样本量增大时趋近于公式 5.6,对其更为有用、但没那么准确的描述是:

---

① 原文为 $(1/2)n(n-k1)$,怀疑有误。——译者注

$$\text{近似的 rho} = 1 - \frac{E_1}{E_2}$$

在这里 $E_1 = \sum_{i=j}\sum (r_{ij\cdot F})^2/df_k$，$E_2 = \sum_{i=j}\sum (r_{ij})^2/[1/2n(n-1)]$。在此 $r_{ij\cdot F}$是 k 个因子的作用被剔除后变量之间的偏相关系数，df 是自由度，它在探索性因子分析中是 $1/2[(n-r)^2-(n+r)]$，但在验证性因子分析中会更大些（剩余偏相关系数不过是一个标准化后的预测的相关系数和观察到的相关系数之间的差）。

另一种概念的经验证实：抽样准确性

传统的统计检验假设抽样的对象是单位（物体或实体），但现实中我们不能忽略它可能涉及一定程度的心理测量抽样——我们所分析的变量几乎总是潜在的更大领域的相关变量的一个子集。因而我们要考虑给定的数据（变量的子集）是否足以进行因子分析。（读者可能还记得映像因子分析和 Alpha 因子分析都假设过这样的心理测量抽样，但这个问题实际上跟任何类型的因子分析都有关。）

一般来说，在其他情形不变的情况下，经验证实度在下列情况下会更大：(1)变量数增加；(2)共同因子数减少；(3)剩余相关系数减少；(4)因子的确定程度提高。前两个条件与因子模型强加在数据上的经验限制的增加直接相关，第三个条件测量的是因子模型在多大程度上解释了观察到的共变关系，最后一个条件界定了各变量的方差被共同因子解释的多少。最后一个条件和抽样准确性的概念直接相关，因为在其他情形不变的情况下，随着变量数的增加和相关系数大小的平均值的增加，总的因子的确定程度也会增加。

Kaiser (1970, 1974)提出了测量抽样准确性的实用指数，他称之为对"抽样准确性的"总的测量。

$$\text{MSA} = \frac{\sum_{j\neq k}\sum r_{jk}^2}{\sum_{j\neq k}\sum r_{jk}^2 + \sum_{j\neq k}\sum q_{jk}^2} \qquad [5.7]$$

在此 r_{jk}是最初的相关系数，q_{jk}是反映像相关矩阵中的一个元素[①]，它由 $Q = SR^{-1}S$ 得到，其中 R^{-1}是相关矩阵的倒数，且 $S = (\mathrm{diag}R^{-1})^{1/2}$。这个指数值域在 0 和 1 之间。实际上，在且只有在相关矩阵的倒数的所有非对角线元素都为 0 的时候，这个指数才为 1，这又意味着在这种情况下所有变量都可以被组内的其他变量准确无误地估计出来。这个测量的阅读指南如下(Kaiser, 1974)：

0.90 或以上，很高；
0.80 或以上，较高；
0.70 或以上，中等；
0.60 或以上，较低；
0.50 或以上，很低；
0.50 以下，无法接受。

Kaiser 声称分析数据的丰富经验显示，MSA 的大小随以下情况而改进：(1)变量数的增加；(2)共同因子数的减少；(3)个案(实体)数的增加；(4)平均相关系数的增大(1970)。

概括一下，在多大程度上一个给定的因子分析模型适合我们研究的问题，数据所提供的经验证据是因情况而异的。研究者应该了解什么条件可以改善因子分析的资讯性。此外，因子分析的初学者应该依据类似于 Kaiser 的 MSA 之类的实用指数来粗略地判断自己的数据是否适合运用这种技术。当然，理论上的正当性是最终决定的基础。

验证性因子分析

任何验证性因子分析的最起码的要求是我们要事先假设共同因子的数量。

① 原文为 rij，qij，有误。——译者注

然而，如果非要说这个假设跟直觉或猜测不同，那是因为它依据的是我们对所考虑的变量性质的理解，同时也依据哪个因子可能在哪些变量上有负载的预期。对这点无论怎么强调都不过分，因为这些因子分析假设的实际形式几乎是无限的。

我们可以把验证性因子分析分成两大类：(1)只涉及一个群体或总体，(2)涉及两个或更多的群体或总体。下面我们先介绍第一类。

一个群体或总体

给定一个群体的协方差矩阵，在验证性因子分析中，我们从假设用于解释观察到的协方差结构的因子结构开始。然后我们评价观察到的数据结构是否“显著地”偏离假设的结构。在极端情况下，假设可能设定：(1)共同因子的数量；(2)因子之间关系的性质——正交或者斜交；(3)各变量因子负载的大小。在另一种极端情况下，假设可能只是设定了潜在的共同因子的数量。当然，很多可能的假设都在这两个极端之间。

形式最简单的验证性因子分析只设定了共同因子的数量，因为它和探索性因子分析区别不大，这里只需要进行简单的解释。对这种假设，我们用正交或斜交因子模型都不重要，要么用显著性检验，要么用其他标准(例如信度系数)来评价初始因子分析结果是否充分就可以了。因此，这里要指出的唯一一个差别是我们根据事先的考虑由一定数量的因子开始，而在探索性分析中我们由一个“保险的”数量开始，如果发现最初猜测(的数量)不充分，就要增加这个数量。需要提醒的是，完全依赖显著性检验是不明智的，除非你愿意接受实质上不那么重要但统计上显著的因子。如果使用了经验认定的方式，我们最好是旋转初始的因子结果以判断产生的结构是否有意义。

另一个极端也很容易说明。如果研究者对因子数量、因子之间关系的性质、因子负载有特定的假设，无论是检查根据假设重新生成的相关矩阵和观察到的相关关系之间是否足够相似，或者是把假设作为目标矩阵，寻找与其能有

最大程度的相似性、且能最大程度重新生成观察到的相关关系的矩阵,都是可能的。在前一种情况下,对假设是否充分的检查取决于对协方差矩阵之间相似性的检验,在后一种情况下,对两个因子结果之间的相似性进行检验以做评估是有必要的。读者可以参考本丛书中的一本来了解更详细的信息(Levine,1977)。一般来说,在实际的因子分析中我们不能预期研究者能有如此明确的假设。然而,如果我们正在比较根据一个数据而来的因子分析结构和根据其他数据而来的结构,我们就有可能作出这样的假设。这种情况可被归纳到我们下一个即将讨论的更一般的方法下。

因为 Sörbom 和 Jöreskog 最新的验证性因子分析程序允许较大程度的灵活性,因此我们会简单介绍程序中一些重要的设定选项(Sörbom & Jöreskog,1976)。我们可以用几种方法设定任何参数。在因子分析中涉及的参数包括因子负载(共同因子的 nr)和因子之间的相关系数 (1/2r(r-1)),这些参数中的任何一个都可以固定为特定值或者允许它自由变动。通常来说,最有用的固定参数的方法是设定某个负载为 0。举例来说,如果固定所有因子相关关系为 0,也就等于设定了正交旋转的结果。另一种设定参数的方法是限定某个参数等于另一个参数。

为了使前面的讨论更加具体,我们在表 5.2 中给出了 3 个设定自由的和固定的参数的例子。在这些例子中,X 代表可自由变化的参数,0 代表固定的参数——固定为 0。我们可以把参数固定为其他值,例如 1.0、0.5,等等,但我们认为研究者也许只知道在哪里负载会高,在哪里负载会低。第一个假设设定一个单因子的结构——在给定变量数后可能的最简单的结构,第二个假设预期有一个一般性因子和两个群体因子,第三个假设设定一个特殊的层级关系。研究者当然也可以在这些模式中引入很多修正。

在进行上面例示内容的同时,我们也要设定因子之间关系的性质。这种设定最常见的形式是:(1)将所有的因子关系设为 0——即正交假设;(2)允许所有相关关系变化——即斜交假设;(3)混合形式,即其中一些假定为正交,其他则允许取任何值。

表 5.2 验证性因子分析设定参数的 3 个例子[a]

变量	例 1			例 2			例 3		
	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
X_1	X	0	0	X	X	0	X	X	X
X_2	X	0	0	X	X	0	X	X	X
X_3	X	0	0	X	X	0	X	X	X
X_4	0	X	0	X	X	0	X	X	0
X_5	0	X	0	X	0	X	X	X	0
X_6	0	X	0	X	0	X	X	0	0
X_7	0	0	X	X	0	X	X	0	0
X_8	0	0	X	X	0	X	X	0	0

注：a. X 代表可自由变化的参数，0 代表固定为 0 的参数。

表 5.3 解决斜交因子模式的固定的和自有变化的参数[a]

变量	因子	
	F_1	F_2
X_1	X	0
X_2	X	0
X_3	X	0
X_4	0	X
X_5	0	X
X_6	0	X
因子之间的相关系数		
	F_1	F_2
F_1	1	
F_2		1

注：a. 0 表示固定的参数，X 表示可以自由变化的参数。在录入协方差矩阵后，因子相关矩阵中的 1 可以认为是固定的。然而，在计算自由度时，我们把这些值看做是一般性设定的一部分，所以我们不把 1 看做是固定的。

表 5.3 给出了验证性因子分析的例子，采用的是表 1.1 的样本数据。如果研究者愿意，可以假设：(1)有两个潜在的共同因子；(2)两个因子可能是相关的；(3)其中一个因子在 X_4，X_5 和 X_6 上的负载为 0，另一个因子在 X_1，X_2 和 X_3 上的负载为 0。

请注意，与探索性因子分析相比，我们在 12(nr)个参数中固定了 6 个参数(因子负载)，但在因子协方差矩阵中允许一个参数自由变化。所以，我们强加了 5 个附加的限制。但不是所有这 5 个限制都会反映在增加了的自由度上。在探索性因子分析中，我们也用了 $\frac{1}{2}r(r-1)$ 个隐含的条件以使某个特定结果唯一。因此，增加了的自由度是 $5-\frac{1}{2}r(r-1)=4$。一般来说，有固定参数值的模型的拟合程度会比没有固定参数值的模型差。但如果

假设的模型是合适的,增加了的自由度不仅弥补了损失的拟合,而且还增进了拟合。

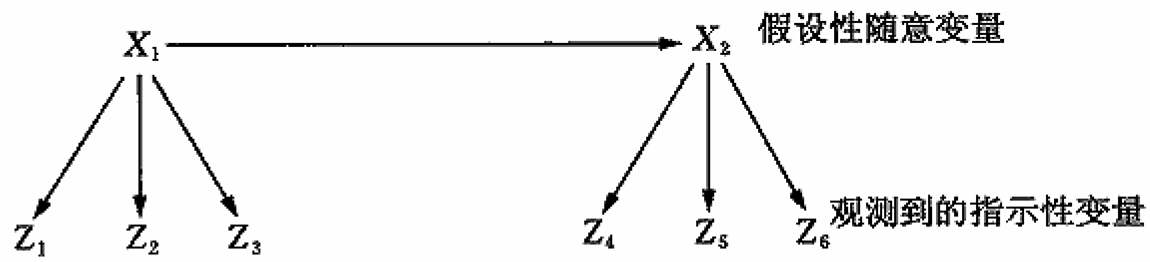


图 5.1　涉及两个假设性变量和指示性变量的因果模型

同时要注意的是,对一个 6 变量的模型运用 3 个共同因子的模型没有多大意义。然而,如果假设的因子结构有足够的限制来获得许多的自由度,把这样一个模型运用在这些数据上是可能的。举个例子:设定 X_1 和 X_2 只在因子 1 上有负载,X_3 和 X_4 只在因子 2 上有负载,X_5 和 X_6 只在因子 3 上有负载。

在此我们提醒一下,本章所例示的原则可以被一般化(超越纯粹的因子分析)到协方差矩阵。尤其是可以结合因子分析和通径分析或回归分析的特点。举例来说,如果我们有一组变量表示一个理论性变量(F_1)的指标,它影响另一个理论性变量(F_2),对这个理论性变量我们也有一组指示性变量,这样的关系系统可以很容易地用验证性因子分析的扩展形式来分析。在这个特定的例子中,模型可以被界定为一个有 2 个相关因子的验证性因子分析,正如图 5.1 所表示的那样。要注意到这样的模型跟表 5.2(样本 1)所界定的结构没什么不同,它们都对因子之间的关系没有限制。在这里我们只是提到了验证性因子分析最简单的扩展形式;有兴趣的读者可以参考 Jöreskog(1970)以及 Sörbom 和 Jöreskog(1976)更完整的展示。

比较因子结构

验证性因子分析的另一种用法是对几个群体比较因子结构。举例来说,我们可以假设黑人政治态度的因子结构和白人的是一样的,或者一个社会的认知结构和另一个社会是一样的。我们也可以界定因子结构的某些方面在不同群体中是一样的,但另一些方面却是不同的。

由 Jöreskog 和 Sörbom 设计的、可以进行模型修改的验证性因子分析程序 COFAMM：Confirmatory Factor Analysis with Model Modifications，可以处理非常一般性的假设。例如它允许对单个群体进行各种因子假设检验——一些参数可以是固定的，或者允许自由变化，或者一些参数可以限制为等于另一些参数。此外，通过使用“限制的”参数，一个群体的参数结构的任何一部分都可以设定为和另一个群体的该部分相同。

作为示例，我们再看一下政治态度的例子，我们感兴趣的是比较白人的因子结构和黑人的因子结构。特定的假设可以有以下形式：(1)白人和黑人都有两个斜交因子；(2)变量 X_1(对学校的资金支持)，X_2(降低失业率的开支)和 X_3(对大企业的控制)在白人和黑人相同的因子上有相同的负载；同样，对两个种族而言，X_4(校车接送计划)和 X_5(工作配额)都在另一个因子上有负载；(3)这两个群体在 X_6(启蒙计划)上的预期负载会不同。在这种情况下，我们可以在单群体的分析中设定白人的参数，所有黑人的参数都限制为和白人相同，除了一个(涉及到 X_6 的)参数。验证性因子分析的详细例子，以及更一般化的“协方差结构”分析都可以在 Jöreskog(1967)的著作中找到。

第 6 章 | 建立因子尺度

检查了因子分析结果后，我们出于两种不同的原因需要建立因子尺度。第一，在发现了数据中的一些潜在维度后，研究者可能想从这些维度而不是各个变量来研究个案。第二，研究者也许在另一个研究中想用一个或多个因子作为变量。实际上，除了心理测量的文献，因子分析似乎更经常被用做为其他研究建立因子尺度的工具，而不是作为研究因子结构本身的工具。在本章中，我们会讨论建立因子尺度的多种方法。会讨论到的方法包括：(1)回归估计；(2)基于理想变量的估计，或者"最小二乘"原则；(3)Bartlett 的最小化误差方差方法；(4)有正交性限制的估计；(5)简单地把有高因子负载的变量相加；(6)建立主成分尺度。这些方法将在建立因子尺度的多种重要背景下讨论。

因子尺度的不确定性

首先，假定我们的数据没有误差，也假定数据是由一个单个共同因子的模型生成的。建立因子尺度的主要目标是以观察到的变量 Xs 为基础决定每个个案在共同因子(F)上的取值。回顾我们迄今讨论的材料，很明显我们可能无法由变量准确地确定共同因子，因为各变量都包含一个特有的成分，且它与变量的共同部分混合在一起无法分离。一般来说，我们能做的最多就是从变量中得到对共同因子的取值的估计。由于这个原因，因此我们说建立因子尺度总有一些不确定性。

为了说明这一点，让我们考虑一个有 3 个变量的单个共同因子的模型。需要特别指出的是，我们假定所有因子负载都相同(或者所有相关系数大小都相

同)。这样的例子展示在图 6.1 的左半部分。在这样的模型中,观察到的变量之间的相关关系等于各因子负载的乘积,因此在本例中它就等于一个因子负载的平方,因为所有的因子负载都相等:

$$r_{ij} = b_i b_j = b_i^2 = b_j^2 = h^2。 \quad [6.1]$$

这个公式也显示出观察到的相关关系等于变量的公因子。

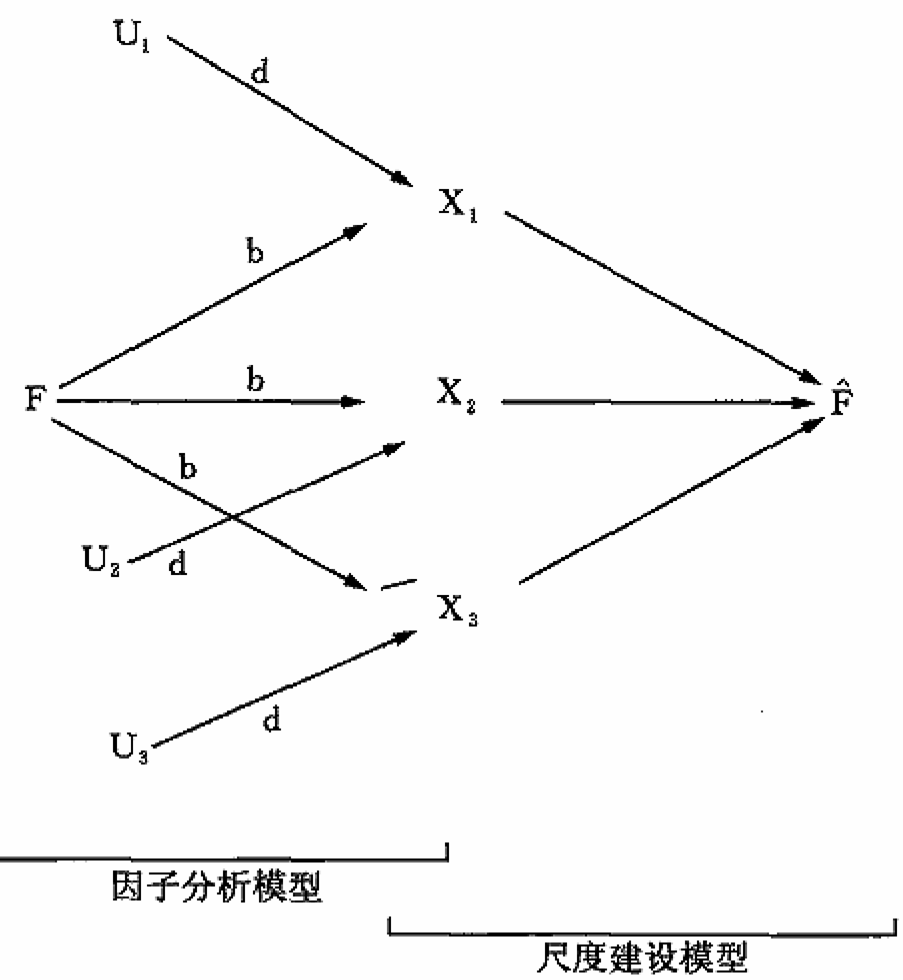


图 6.1 表示因子和因子得分的路径模型

我们继而通过合并观察到的 Xs 来建立一个指标(如果你喜欢,也可以称之为因子尺度)。因为每个变量在共同因子上有相同的负载,因此这种情况下把它们相同的权重再相加是合适的。我们得到的指标可以用以下公式表示

$$\hat{F} = X_1 + X_2 + X_3,$$

这其中包括图 6.1 右半边表示的因果指向。尤其要注意到指标 $\hat{F}$ 有 4 个最终的来源变量——共同因子 F 和 3 个独有的因子,U_1, U_2 和 U_3。因此,由于独有因子的存在,F 和 $\hat{F}$ 之间的相关不会是完美的。下面我们研究潜在的共同因

子和因子尺度之间的相关程度,即尺度的信度。

因子尺度的信度

尺度($\hat{F}$)的方差很容易通过应用第一卷书第2章所用的期望运算导出。得到的 $\hat{F}$ 的方差用 Xs 的方差表示则为:

$$\begin{aligned} var(\hat{F}) &= var(X_1) + var(X_2) + var(X_3) \\ &\quad + 2[Cov(X_1, X_2) + Cov(X_1, X_3) + Cov(X_2, X_3)]。 \end{aligned} \tag{6.2}$$

该简化是由于在这个例子中权重为1。这个公式可以通过假定各变量的方差为1来继续简化,这样协方差不仅和相关系数相等,且互相相等:

$$\begin{aligned} Var(\hat{F}) &= n + 2[r_{12} + r_{13} + r_{23}] \\ &= n + n(n-1)r \end{aligned} \tag{6.3}$$

(这样实际上就相当于把 Xs 相关矩阵中的所有元素相加。)

$$\begin{aligned} &= n[1 + (n-1)r] \\ &= n[1 + (n-1)h^2] \end{aligned}$$

因为 $r_{12} = r_{13} = r_{23} = r = h_i^2$ (参见公式6.1)。

然而,$\hat{F}$ 中的一些方差由独有因子所贡献,它们的贡献为 $\sum d_i^2 = \sum (1 - h_i^2)$,它可以继续被简化为 $n(1 - h^2)$。因为在这个例子中所有公因子都被假定为相等,因此,$\hat{F}$ 的方差中被公因子解释的部分为

$$\begin{aligned} r^2_{(F,\hat{F})} &= \frac{Var(\hat{F}) - n(1-h^2)}{var(\hat{F})} \\ &= \frac{n[1+(n-1)h^2] - n(1-h^2)}{n[1+(n-1)h^2]} \\ &= \frac{nh^2}{1+(n-1)h^2} \\ &= \frac{nr}{1+(n-1)r}, \end{aligned} \tag{6.4}$$

它等同于 Spearman-Brown 的信度公式，它是 Cronbach Alpha 系数的特殊情况(Cronbach, 1951; Lord & Novick, 1968)。(读者应该还记得，在这种情况下 h^2 可以被相关系数 r 代替。)

为了让读者熟悉因子尺度的不确定程度或者说预期“信度”，我们通过表 6.1 给出了公因子和变量数的常见组合的“信度”值。我们注意到，当公因子的值(因子负载或者相关系数)固定，随着变量数的增加，信度就增加。同时也注意到，即使有相同的较高的因子负载(例如 0.8)，但如果我们只有很少的几个变量，信度也较小。

我们也应该注意到，在建立尺度的过程中，通常会标准化 $\hat{F}$，使之均值为 0，方差为 1。这样的标准化很直接，且可以被纳入到权重里，但它的价值只是修饰。

表 6.1 不同统一因子负载值和变量数的预期信度
(因子和它的平方后的尺度的相关系数)[a]

因子负载	0.4	0.5	0.6	0.7	0.8	0.9
公因子(h^2)或者变量之间的相关系数	0.16	0.25	0.36	0.49	0.64	0.81
变量数						
2	0.276	0.400	0.529	0.658	0.780	0.895
3	0.364	0.500	0.628	0.742	0.842	0.927
4	0.432	0.571	0.692	0.794	0.877	0.945
6	0.533	0.667	0.771	0.852	0.914	0.962
8	0.604	0.727	0.818	0.885	0.934	0.972
12	0.696	0.800	0.871	0.920	0.955	0.981
20	0.792	0.870	0.918	0.951	0.973	0.988

注：a. 信度(α)的公式 $=\dfrac{n(r)}{1+(n-1)r}=\dfrac{n(h^2)}{1+(n-1)h^2}$。

因子负载不相等的情况

迄今为止我们不仅假定因子负载是统一的，而且假定模型是没有误差的。我们现在研究在更复杂的情况下会发生什么。在一个共同因子模型中因子负载不完全统一的情况下，将会导出一个有不同相关系数的矩阵。如果我们简单地通过把观察到的变量加总而建立一个因子指标，得到的尺度的信度为：

$$\text{Cronbach 的 alpha 系数} = \frac{\text{简化过的相关矩阵中的元素之和}}{\text{相关矩阵中的元素之和}}$$

$$= \frac{Var(\hat{F}) - \sum d_i^2}{Var(\hat{F})} = \frac{Var(\hat{F}) - \sum(1-h_i^2)}{Var(\hat{F})} \qquad [6.5]$$

如果所有公因子都一致,它就等同于公式 6.4。一般来说,给定相同数量的平均公因子(或者平均相关系数),负载统一情况下的信度比不统一时要高。因此,表 6.1 为多种组合的负载很好地展示了信度的上限,这些负载是表中给出的负载的平均值。

然而,更重要的问题是给定不同的因子负载,我们应不应该在建立因子尺度时给予相同的权重。让我们来考虑一个比较极端的例子,其中的一个公因子为 1,即一个观察到的变量和潜在的因子相连不可分。在这种情况下,我们可以用该特定变量来描述潜在的因子,忽略其他的变量;增加其他公因子小于 1 的变量只会歪曲了尺度。

因此,一般来说,当因子负载不统一时,简单地通过把所有变量加总来建立因子尺度是不合适的。如果一个单个共同因子的模型和数据完全吻合,正如这里假定的那样,最优的解决方案相对简单;给予各变量的权重可以由以下得到

$$B'(R^{-1}) \qquad [6.6]$$

(此处 B 是因子负载的向量,R 是 Xs 的相关矩阵)

它等同于把因子回归到变量上得到的回归权重。在此我们最大化 F 和 $\hat{F}$ 之间的相关系数,它的平方可以由下列公式得到:

$$\text{一般化的信度} = \frac{var(\hat{F}) - \sum(1-h^2)w_i^2}{var(\hat{F})} \qquad [6.7]$$

在此 w_i是公式 6.6 中给出的各变量的回归权重,所建立的尺度的总方差由下列公式得到

$$var(\hat{F}) = \sum_i \sum_j w_i w_j r_{ij} \qquad [6.8]$$

它等于把调整后的相关矩阵的所有元素都加总,各元素都乘以相应的权重 w_i和

w_j(调整后矩阵在它的主对角线上会包含给定变量的各个权重)。因为这个值等于多元 R^2,它不会小于最大的公因子。因此,如果一个变量跟潜在的公因子完全一样,这个变量会得到所有的权重,而其他的变量会被忽略。

在建立指标的时候,另一个重点是当我们用不同的权重时,让一个变量有较高的负载通常比让很多变量有中等的负载要重要。同时要记住,尺度的信度至少会跟最高的负载的平方相等。

抽样变异性和模型拟合的不同标准

迄今为止,我们讨论了理想的情况,即单因子模型和数据完全吻合,且没有抽样变异性(当然,这暗示着我们假定潜在的模型完全可以被识别)。当在数据中引入抽样变异性时,我们在样本中观察到的关系就不会跟总体中的真实情况完全一致。即使一个单个共同因子的模型在总体中跟数据完全吻合,这样的模型也不可能完全解释在数据中观察到的样本关系。这个问题要求研究者用一个标准来评价尺度和潜在的因子之间的拟合程度。在文献中有以下三种标准:

回归的方法

第一个标准是找一个因子尺度($\hat{F}$),这个尺度($\hat{F}$)和潜在的共同因子(F)之间的相关系数最大。或者换言之,这个标准就是最小化两者之间的离差的平方,即最小化 $\sum(F-\hat{F})^2$。为了符合这个标准,我们要运用回归的方法。这种方法之所以可能,是因为因子分析给我们提供了因子负载,它们等同于因子(在建立尺度时预测)和观察到的变量(被用做预测变量)之间的相关系数,以及预测变量之间的相关系数,它们只是 Xs 之间观察到的相关系数。这两组相关系数是我们在解一般回归方程时所需要的所有信息。预期得分由以下公式得到

$$\hat{F} = XR^{-1}B \qquad [6.9]$$

此处 B 是因子负载的矩阵,Xs 是观察到的变量,R 是 Xs 的相关矩阵。注意到前面我们已经给出了加权系数(公式 6.6)。唯一的不同是在公式 6.9 中我们可

以用预测的相关系数(BB′)来代替实际的相关系数,这样做不会改变原则,因为在无误差的总体模型中这两者是等同的。在现在的情境下,重新生成的相关系数一般来说跟观察到的相关系数是不同的。尺度的预测信度同样可以由公式6.7得到。

最小二乘法标准

在单个共同因子的模型中,每一个变量都被看做是共同因子和独有因子的加权和:

$$X_j = b_j F + d_j u_j$$

接着我们考虑在因子尺度中用预测的 $\hat{F}$ 代替 F。这个标准是构建一个 $\hat{F}$ 使得以下的平方和最小。

$$\text{最小化} \sum_i \sum_j (X_{ij} - b_j \hat{F})^2 \qquad [6.10]$$

该标准导出的权重为

$$\hat{F} = X(BB')^{-1}B \qquad [6.11]$$

注意到它和公式6.9之间唯一重要的不同之处在于,我们用重新生成的相关关系(BB′)代替了观察到的相关关系(R)。因此,在我们处理总体的变量,且它们和单因子模型完全吻合时,这两个标准会得到相同的指标。然而,当样本的相关关系和总体的相关关系趋异时,这两者也会趋异。

巴特利特标准

第三个标准是在研究拟合程度时考虑抽样变异性。如果我们把独有的方差看做半误差的方差,正如在最大似然因子分析法中的那样,给包含更多随机误差的变量更少权重,而给包含更少误差的变量更多权重是合理的。因此,这里使用的准则是在对每个元素用误差方差的倒数加权后,最小化公式6.10中给出的平方和。这个标准是最小化:

$$\sum_i \sum_j (X_{ij} - b_j \hat{F})^2 / d_i^2 \qquad [6.12]$$

结果是包含公因子较小的变量被给予更小的权重。因此,在因子负载不统一的情况下,这个标准得到的尺度跟前面两种标准所得到的会不一样。导出因子尺度的公式(6.13)看起来很复杂,但其背后的原则并不难以理解:

$$\hat{F} = XU^{-2}B(B'U^{-2}B)^{-1} \qquad [6.13]$$

此处 U^2是独有方差的对角线矩阵。在公式中 U^{-2}可以被视为前面描述过的一种加权方法。

多个共同因子和更复杂的情况

让我们假设有超过两个共同因子,把情况设置得更复杂些。前一节中讨论过的三个标准可以被一般化到多变量的情况下,也可以被一般化到正交或者斜交因子分析法中。此外,我们讨论过的单个因子的情况,在考虑多变量和任何特定共同因子时也成立。然而,我们建立的尺度一般来说不可能和相应的潜在因子完全相关,这会在多个因子的情境下引出两个附加的问题:(1)如果潜在的因子之间是正交的,这些不完美的尺度之间会不会也是正交的;(2)各个尺度是不是只和它试图测量的因子相关,而不与其他变量相关,如果因子尺度和其他潜在的因子之间的偏相关在控制了它测量的因子之后为 0,那么该因子尺度是否就被称做是单一的? 一般来说,任何一种建立尺度的方法,这些要求都不能同时被满足。即使潜在的因子之间被假定为正交的,因子尺度之间也会互相相关;此外,如果我们假定的是斜交模型,因子尺度之间的相关将不能正确地反映因子之间潜在的相关。最后,一个因子的尺度会和其他潜在的因子相关。

但在一种特殊情况下这些要求都能满足:(1)因子分析性模型和数据完全吻合,且没有任何抽样或测量误差;(2)每个变量仅在一个因子上有负载。如果这两个条件都满足,我们可以分别考虑每个因子或维度,这就回到了我们前面讨论过的无误差数据下的单个共同因子模型的情况。我们之前提到过,在那种情况下,在选择建立尺度的方法上没有任何模棱两可的地方——我们研究过的

所有标准都会得到一样的尺度。但遗憾的是,这种理想情况在实际中不可能出现。

然而,在另一种情况下有一些建立尺度的方法可以满足尺度的正交性和单一性要求。如果最初未旋转的因子是通过最大似然法(或者正则方法)得到的,且如果这些尺度是通过回归或者 Bortlett 方法建立的,那么这些未旋转的因子的因子尺度将会是正交的和单一的。然而,这只提供了些许安慰,因为我们不能期待潜在的因子模型是正交的。此外,在正交旋转后回归方法不能满足任何一个要求,而 Bortlett 的方法只能满足单一性的要求。因此我们讨论过的任何一种方法都不可能得到正交的尺度。

这些结果促使我们考虑 Anderson 和 Rubin (1956)提出的第四种标准。Anderson-Rubin 标准是 Bortlett 标准的修正,它最小化 Bortlett 标准所用的加权平方和,其限制是建立的尺度要互相正交。因此,不管因子是否被旋转过,只要它们是基于正交旋转的方法,这个标准就能够产生互不相关的因子尺度。然而,这些尺度对旋转后的结果来说不是单一的,即使初始的旋转结果是基于最大似然法。

如何选择

在这些方法中进行选择时,研究者必须考虑这些方法内在的特点,同时也要考虑因子分析以外的要求。以下我们对这些方法内在的特性做一些综述。就潜在的因子和它相应的尺度之间的相关关系来说,回归法要比 Bortlett 方法更好,但 Bortlett 方法又要比最小二乘法好。就单一性的要求而言,Bortlett 方法最好,但就正交性的要求来说,Anderson-Rubin 的标准更理想。然而,考虑到在大多数研究情境下,研究者不大可能会坚持潜在因子之间的正交性,因此实际的选择就在回归法和 Bortlett 方法之间。

我们也要讨论其他一些因素,这些因素有的把选择变得更复杂,有的把选择简单化。首先,用不同方法得到的尺度之间通常有很高的相关关系,因此,对任何研究来说,选择都应该是基于学术考虑的。建立尺度的一种方法和另一种方法一样,都要能够满足要求(Horn, 1965; Alwin, 1973)。第二,选择也依赖于

要研究的特定问题。Tucker(1971)指出,当因子尺度和外部变量相关时,一些方法在某些特定类型的研究中更合适。更具体地说,他证明了通过回归法得到的尺度无法让我们正确地估计假设的因子和外部变量之间的相关系数,但其他的方法则能让我们做到这一点。另一方面,如果使用因子尺度的主要目的是为了把它作为外部变量的预测变量,那么基于回归法得到的尺度也许会更好。

然而,要注意的是,我们迄今为止的讨论都是基于一个假设,即因子模型在总体中和数据完全吻合,所以模型和数据之间的任何差别都被假定为随机抽样误差的结果。但是,如果我们预期的这种完全吻合不出现,或者我们只是把因子分析用做在数据中把变量分成主要大类的探索性方法,此时我们之前探讨的所有细节都成了次要问题,因子分析以外的考虑则更重要。

基于因子的尺度

基于两种完全不同的原因,我们可能考虑只运用因子分析中得到的一些信息来建立尺度,而不是依赖于我们迄今为止讨论的因子尺度。第一,我们可以接受因子分析性模型在总体中与数据完全吻合的假设,但同时也认为在因子分析结果中得到的一些值会受到抽样误差的影响。在此我们可以忽略因子负载中的特定变化,只把一类信息看做相关的:一个变量要么在给定的因子上有负载,要么没有负载。第二,尺度可以通过对有足量负载的变量加总、忽略只有少量负载的其余变量而得到。通过这种方法建立的尺度不再是因子尺度,而只是基于因子的尺度。这样建立尺度背后的原因是:(1)即使在总体中一些变量的因子负载为 0,它们在特定的样本中也不为 0;(2)即使在总体中因子负载是统一的,它们在样本中也不会如此。这种情景下的常规做法是把因子负载小于 0.3 看做是不足量的负载。

这种建立尺度的方法是否合理,取决于它背后特定的假设在多大程度上合适。理想状况下,我们可以用验证性因子分析来检验这些假设。然而,如果这些“简单的负载模式”被验证性分析所支持,那么它就不再是一个基于因子的尺度,而是一个合理的因子尺度。但是,实际上即使我们进行这样的检验,而且它

们确实也表现出统计上显著的偏差,但这也只是程度的问题,而且与简单负载之间的少量偏差仍然可以被忽略,下面我们说明具体原因。

解释建立简单指标做法的合理性的基础则完全不同。(我们假定这类尺度方法是最简单的,但这并不是使用它最重要的理由。)通常我们不会预期因子分析性模型和数据完全吻合,原因有以下两点:(1)变量的非随机测量误差;(2)没有设定的,以及在概念上和我们感兴趣的领域不相关的次要因子也许能够解释观察到的部分相关关系,继而影响我们得到的权重。因此,我们不能只看一个给定的因子分析结果的取值。保守的做法是把因子分析中得到的结构只看做是建议性的,只表示数据中的集群,而没有其他。换言之,我们一旦认为所得到的特定数量(的因子)包含了相当程度的"噪音",聪明的做法就是忽略次要的差别和偏离。

表 6.2 验证性因子分析的结果,使用的是表 1.1 上三角区的相关矩阵和表 3.2 设定的模型*

变量	因子		公因子
	F_1	F_2	h^2
X_1	0.792	0	0.624
X_2	0.756	0	0.571
X_3	0.633	0	0.501
X_4	0	0.577	0.333
X_5	0	0.669	0.448
X_6	0	0.635	0.404
$r_{F_1F_2}=0.501$			

$\chi^2=4.6534$
df = 8
可能性 = 0.7939

注:* 这些结果通过 LISREL Ⅲ 得到,而不是通过 COFAMM 得到。

也许有人会反对在建立这种尺度时所采取的漫不经心的态度。也有人可能会顾虑简单地把观察到的变量进行组合也许不是最优的方案,因为给予不同的权重会增加尺度和观察到的变量之间的总的相关(也就是说,他们顾虑的是把简单的加总作为代表原始变量中包含的信息这一方法是否高效)。即使考虑到这一点,简单的加权仍然是合理的,因为权重的细微变化不会导致尺度(变量的某种组合)和所有变量的多重相关关系有多大的变化(Wang & Stanley, 1970; Wainer, 1976)。第 3 章提到的一个问题要谨慎处理。如果因子分析性模型被认为如实地反映了数据,那么忽视很高的因子负载——例如 0.9——且给予它们与低负载的项目相同的权重,将会有不良后果。总而言之,我们认为因子尺度和基于因子的尺度在实际研究中都有它们合理的位置。

成分得分

最后，我们想评论一下由主成分分析法产生的尺度。正如之前提到(强调)的，主成分分析法背后的原则与因子分析不同。因此，它们不能相互代替。但在实际研究的更大背景中，它们都有合理的应用。在一些情况下成分得分可能比因子尺度更合适。尤其是如果目的只是对原始数据中包含的信息进行简单的总结，而不需要因子分析性的假定，那么使用成分得分绝对要比因子尺度好。这也是为什么对成分尺度进行评论很重要的原因，即使只是简单的评论。

正如前面提到的，主成分法不过是对原始变量进行准确的数学转换。因此，用原始变量的组合来确切地表示成分是可能的，我们可以称其为成分得分，而不是尺度或估计。这个得分从各种原始变量的组合中得到，这些原始变量的权重跟他们的成分(因子)负载成比例。

$$\text{成分得分} = \sum_j [(b_{ij}/\lambda_i)X_j] \qquad [6.14]$$

在此 b_{ij} 是第 j 个变量在第 i 个成分上的负载，λ_i 是相应的特征值。公式中除以特征值只有修饰作用，因为它只是保证最后得到的指标的方差为 1。

第 7 章 | 对常见问题的简单回答

与变量的性质以及它们的测量有关的问题

1. *因子分析要求测量哪个层次的变量?*

因子分析要求变量的测量至少为定距层次(Stevens, 1946)。在因子分析中要把相关矩阵或协方差矩阵作为基本数据就暗示了这个要求。此外,把变量设定为潜在因子的加权和,以及把因子尺度的建立作为观察到的变量的加权和,在定序或定类变量的情况下都无法确定。

2. *用 Kendall 的 tau,以及 Goodman 和 Kruskal 的 gamma 作为相关关系的测量代替一般的相关系数是否合适?*

不合适。正如上面提到的,对定序变量无法进行加法运算,因此因子分析性模型不能整合定序变量。我们可以在因子分析中把这样的矩阵仅仅只作为探索用途,但我们无法给予这些结果统计上的解释。(有一些非度量尺度法被设计出来专门处理非度量变量。)

3. *基于以上的回答,在变量的矩阵基础不明确时,研究者是否应一直避免运用因子分析?*

不一定。许多变量,例如对态度或观点的测量,以及许多测试成就的项目都没有明确的矩阵基础。但是,我们一般假定很多“定序变量”可以给予数值而不扭曲其潜在属性。对这个问题的最终回答取决于两个考虑:(1)任意给定的值在多大程度上反映了潜在的距离;(2)调整尺度的扭曲程度多大程度上导致了相关关系(将作为因子分析的基本数据)的扭曲程度。幸运的是,相关系数在被扭曲为定序层次的情况下仍较为稳定(Labovitz, 1967, 1970;

Kim, 1975)。因此,只要我们认为对定序类别给予数值所造成的扭曲不是很严重,那么把定序变量作为度量变量仍是合理的。然而,我们应该注意到在因子分析性结果中由于非随机测量误差导致的可能的扭曲,即使这些扭曲是轻微的。

4. 二分变量如何?有些人认为把因子分析应用到二分变量上没有问题,原因有二:(1)因为给二分类别赋值并不真的要求有测量假设;(2)因为前述观点,ϕ和皮尔逊相关系数等同,而它是因子分析中对相关关系的合适测量。因此,把因子分析应用到ϕ的矩阵中难道不合适吗?

不合适。首先,我们不能在因子分析性模型中表达二分变量。更具体地说,回顾 Kim/Mueller, University Paper 07—013 的第 2 节,在因子分析中,每个变量都被假定为是至少两个潜在因子(一个共同因子和一个独有因子)的加权和。即使这些潜在的因子也有两个值,正如在 University Paper 07—013 中的表 1 所展示的那样(在真正的因子模型中这种情况很少出现),观察到的变量的值至少包含了 4 个不同值,这明显不符合二分变量的情况。因此,无论如何在二分数据中运用因子分析法都是不合适的,除非只是为了探索。以下 3 个问题也是相关的。

5. 前述回答暗含的意思让人担忧。因为我们一般把潜在的因子看做是可能连续的,所以我们预期的变量包含许多类别。然而,我们处理的大多数变量通常只有有限的类别——是或否,同意或不同意,或者最多有强烈同意,同意,中立,不同意,强烈不同意,等等。这是不是真的就意味着我们把因子分析用在了不符合因子分析模型的数据上了?

从某种意义上说,是的。严格来说,只有少量类别的变量是不符合因子分析性模型的。如果我们把观察到的变量看做代表了粗略的测量或对相邻值的分组,问题就不再是数据是否内在的符合因子分析模型,而是非随机测量误差是否以及在多大程度上扭曲了因子分析的结果。对数值进行分组当然影响了相关系数,但其程度取决于分组的粗糙程度、分布的形状,等等。然而,即使在严重的测量扭曲情况下,把因子分析作为探索性的工具也有一些令人鼓舞的看法(参见下一个问题)。

6. 在什么情况下，我们可以把因子分析运用到包含二分变量或只有少量类别变量的数据上？

一般来说，类别的数量越多，扭曲的程度越小。即使对二分变量，如果因子分析只用做寻找一般变量丛的工具，且如果变量之间潜在的相关系数被认为是中等的——例如小于0.6或0.7，用phi也是合理的。原因是连续变量的二分化弱化了相关关系，这个弱化也受到切点的影响。然而，当潜在的相关关系不是很高时，切点的变化对相关系数的影响可以忽略。因此，分组一般弱化了相关关系，但不会影响数据中集群的结构——因为因子分析只依赖于相关关系的相对程度。如果研究者的目的是寻找集群模式，使用因子分析则是合理的(Kim, Nie & Verba, 1977)。

7. 如果因为切点导致的扭曲比由于分组导致的（相关关系的）弱化更严重，为什么不用如 ϕ/ϕ_{max} 或者 R/R_{max} 代替 ϕ 或者 r 进行调整？

这个修正只有在潜在的分布形状有特定的形式时（可能性很低）(Carrol, 1961)，或者当潜在的连续变量之间完全相关时，才是合适的。如果是完全相关，进行因子分析就没有意义。因此，使用这样的修正是自相矛盾的(Kim et al., 1977)。

8. 有没有可以更直接地解决这些测量问题的方法？

文献中提出了两种方法。这两种方法都假设二分或多分变量是通过分组或门槛作用而得到，且这些变量是潜在的连续变量的指标变量，因子分析实际上是运用到这些潜在的连续变量上。因此，为了解决因子结构的问题，我们要找出潜在的变量之间的相关关系。其中一种做法是用四分相关系数代替phi。这个方法只是探索性的，因为四分相关经常出问题，且其相关矩阵不一定是格拉姆式的(Bock & Lieberman, 1970)。另一个方法是直接处理潜在的多变量的分布，而不是基于双变量的表格来计算四分相关系数。这是个有潜力的突破口，但即使使用现代计算机，其计算量也很大(Christoffersson, 1975)。

与使用相关或协方差矩阵有关的问题

1. *用协方差矩阵和用相关矩阵有什么不同吗?*

这取决于:(1)变量是否有可相比的矩阵;(2)使用的是哪种抽取因子的方法;(3)我们是否要比较因子结构。如果我们只考虑一个群体(或样本),并且使用的是无尺度的抽取方法,例如最大似然法、Alpha 因子分析法或者映像分析,且我们的目的是确定潜在的维度,那么使用哪种矩阵都没有多大区别。然而,如果我们用协方差矩阵,且尺度变化很大,尺度因子会把对结果的解释复杂化。因此,如果变量之间的方差有差别,且尺度变化很大,例如一个变量的单位是美元,另一个变量的单位是年龄或受教育年限,第三个变量是一个 5 分的 Likert 特量表。此外,出于实际操作的考虑,我们推荐相关矩阵。一些计算机程序不接受协方差矩阵,而且文献中的许多例子都是基于相关矩阵。

2. *什么时候用协方差矩阵更好?*

如果考虑要比较不同群体的因子结构,用协方差矩阵更好。原因是相关矩阵是通过按照样本特有的标准——例如样本均值和方差——来调整变量而得到。由于这个原因,即使理论上认为因子分析涉及不变的因素,我们也不能预期它们在群体(样本)之间不变,因为在计算相关系数时,测量尺度在不同群体间被重新标准化了(Kim & Mueller, 1979,对一般性因果分析中标准化变量的后果的解释和讨论;也可参考 Sörbom & Jöreskog, 1976: 90)。

3. *如果目的是在不同群体间比较因子结构,而变量的测量尺度差别较大,我们应该怎么做?*

一种方法是用一般的标准,例如把合并后的群体作为参照群体的均值或者方差,来标准化变量。这样就可以计算各群体的方差—协方差矩阵。这跟使用群体各自的相关矩阵不同,因为那么做意味着用群体(或样本)各自的标准来转化各个群体的变量。

与显著性检验和因子分析结果稳定性相关的问题

1. 如果运用最大似然法和相关的显著性检验,要求的最小样本量是多大?

样本量越大,χ^2 的估计越好。Lawley 和 Maxwell(1971)建议样本量至少比考虑的变量数多 51,这时检验才是合适的。也就是说,$N-n-1\geqslant 50$,在这里 N 是样本量,n 是变量数。当然,这只是一般性的经验原则。

2. 各假设性的因子应该有多少个对应的变量?

Thurstone 建议每个因子至少有 3 个变量,但如果运用验证性因子分析就没必要满足这个要求。一般来说,研究者大多同意变量数至少为因子数的两倍。

3. 变量的多元正态分布假定总是必要的吗?

因子分析性模型本身并不要求这样的假定。例如,即使用二分因子,我们也可能建立一个因子分析性模型。然而,最大似然法和相关的显著性检验要求这个假定。但一般来说,我们对违反这个假定的后果并没有清晰的理解。

其他各种统计问题

1. 因子负载的正负号含义是什么?

正负号本身没有内在的含义,它也不应该用来估计变量和因子之间的关系大小。但是,相对于其他变量的正负号而言,对给定因子的变量的正负号有其特有含义,不同的符号意味着变量和因子之间的关系是相反的。因此,最好在因子分析前把变量编码成相同的方向。

2. 与旋转后的因子相关的特征值的含义是什么?一个给定的旋转后的因子所解释的方差比例意味着什么?

与未旋转的因子相关的特征值跟旋转后的因子的特征值的意义不同,除了特征值的和相同。在最初的因子分析结果里,特征值的降低程度告诉我们各因子的相对重要性。旋转后的结果则不同。一旦我们通过旋转分离了

不同的维度，数据中的方差作为整体有多少被各因子所解释就不那么重要了。

3. 为了得到“更高层次的”因子分析结果，用因子尺度之间的关系来进行因子分析合适吗？

不合适。因子尺度之间的相关跟潜在的因子之间的相关不同。我们应该用由独有因子结果产生的相关矩阵数据进行更高层次的因子分析。

4. 如果正交的因子结构跟数据吻合，我们可以声称潜在的因子结构是正交的吗？

不可以。正交性是由研究者强加的。然而，如果我们应用斜交旋转法却得到一个正交结构，或者在呈现出来的图像上我们看到变量成直角集群，这时我们可以声称潜在的结构是正交的。

5. 我们可以引入一些变量，而其中一些是另一些变量的原因吗？也就是说，有没有必要保证引入的所有变量在因果顺序中处于同一层次？

一般来说，引入的变量不应该是另一些变量的原因。因子模型假定所有观察到的变量都由潜在的因子导致。然而，出于其他目的，有经验的用户可以把因子分析应用到变量的因果系统中(Stinchcombe, 1971)。

与书、期刊和计算机程序有关的问题

1. 有没有什么因子分析的书或者文章是初学者能读懂和理解的？

实际上没有。大多数都要求有一定的技术背景。不过以下的材料相对来说要容易些：Rummel (1967), Schuessler (1971), Cattell (1952), Comrey (1973), Fruchter(1954)。

2. 哪些书是认真的读者应该考虑研究的“更深入的”书？

Harman(1976), Mulaik(1972), Lawley 和 Maxwell(1971)。

3. 周期性地发表因子分析文章的主要期刊有哪些？

Psychometrika; *British Journal of Mathematical and Statistical Psychology*; *Educational and Psychological Measurement*。

4. 哪些一般性的统计包含因子分析程序?

SPSS, OSIRIS, SAS, BMD。

5. 我们应该知道哪些更专门的处理因子分析的程序?

Kaiser's-Little Jiffy, Mark IV; Sörbom and Jöreskog, COFAMM。

6. 主要的模拟研究报告可以从哪里获取?

Tucker, Koopman & Linn(1969); Browne(1968); Linn(1968); Hakstian (1971); Hakstian & Abell(1974)。

注释

[1] 当潜在的结构很复杂，例如像 Thurstone 的盒子问题那样，通常很难纯粹基于分析性的原则，准确地从协方差矩阵中重制出潜在的模式。我们也许需要拟合超平面和可视化旋转的帮助。

参考文献

ALWIN, D. F. (1973) The use of factor analysis in the construction of linear composites in social research. *Sociological Methods and Research 2*: 191—214.

ANDERSON, T. W. and H. RUBIN(1956) Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 5*: 111—150.

ASHER, H. (1976) Causal Modeling. Sage University Papers on Quantitative Applications in the Social Sciences, 07—003. Beverly Hills and London: Sage Pub.

BMDP-77: *Biomedical Computer Programs* (*P-Series*). W. J. Dixon, Series Editor, M. B. Brown, Editor 1977 edition. Los Angeles: Univ. of California Press, 1977.

BARGMANN, R. E. (1957) A Study of Independence and Dependence in Multivariate Normal Analysis. *Mimeo Series No. 186*. Chapel Hill, N. C.: Institute of Statistics.

BARTLETT, M. S. (1937) The statistical conception of method factors. *British Journal of Psychology 28*: 97—104.

BOCK, R. D. and R. E. BARGMANN(1966) Analysis of covariance structure. *Psychometrika 31*: 507—534.

BOCK, R. D. and M. LIEBERMAN(1970) Fitting a response model for N dichotomously scored items. *Psychometrika 26*: 347—372.

BOCK, R. D. and A. C. PETERSON(1975) A multivariate correction for attenuation. *Biometrika 62*: 673—678.

BROWNE, M. W. (1968) A comparison of factor analytic techniques. *Psychometrika 33*: 267—334.

COFAMM: Confirmatory Factory Analysis with Model Modification User's Guide. Sörbom, D. and Jöreskog, K. G. Chicago: National Educational Resources, Inc., 1976.

CARROLL, J. B. (1953) Approximating simple structure in factor analysis. *Psychometrika 18*: 23—38.

——(1961) The nature of data, or how to choose a correlation coefficient. *Psychometrika 26*: 347—372.

CATTELL, R. B. (1952) Factor Analysis. New York: Harper and Bros.

——(1965) Factor analysis: an introduction to essentials. (Ⅰ) the purpose and underlying models, (Ⅱ) the role of factor analysis in research. *Biometrics 21*: 190—215, 405—435.

——(1966) *Handbook of Multivariate Experimental Psychology*. Chicago: Rand McNally.

——and J. L. MUERLE(1960) The "maxplane" program for factor rotation to oblique simple structure. *Educational and Psychological Measurement 20*: 269—290.

CHRISTOFFERSSON, A. (1975) Factor analysis of dichotomized variables. *Psychometrika 40*: 5—32.

COMREY, A. L. (1973) *A First Course in Factor Analysis*. New York: Academic Press.

CRONBACH, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika 16*: 297—334.

DUNCAN, O. D. (1966) Path analysis: sociological examples. *American Journal of Sociology 72*:

1—16.

EBER, H. W. (1966) Toward oblique simple structure maxplane. *Multivariate Behavioral Research 1*;112—125.

FRUCHTER, B. (1954) *Introduction to Factor Analysis*. New York: Van Nostrand.

GREEN, B. F., Jr. (1976) On the factor score controversy. *Psychometrika 41*;263—266.

GUILFORD, J. P. (1977) The invariance problem in factor analysis. *Educational and Psychological Measurement 37*;11—19.

GUTTMAN, L. (1953) Image theory for the structure of quantitative variates. *Psychometrika 18*;227—296.

——(1954) Some necessary conditions for common factor analysis. *Psychometrika 19*;149—161.

HAKSTIAN, A. R. (1971) A comparative evaluation of several prominent methods of oblique factor transformation. *Psychometrika 36*;175—193.

——and R. A. ABELL(1974) A further comparison of oblique factor transformation methods. *Psychometrika 39*;429—444.

HARMAN, H. H. (1976) *Modern Factor Analysis*. Chicago: University of Chicago Press.

——(in press) Minres method of factor analysis, in K. Enstein, A. Ralston, and H. S. Wilf(eds.) *Statistical Methods for Digital Computers*. New York: John Wiley.

——and W. H. JONES(1966) Factor analysis by minimizing residuals(Minres). *Psychometrika 31*;351—368.

HARMAN, H. H. and Y. FUKUDA(1966) Resolution of the Heywood case in the Minres solution. *Psychometrika 31*;563—571.

HARRIS, C. W. (1962) Some Rao-Guttman relationships. *Psychometrika 27*;247—263.

——(1967) On factors and factor scores. *Psychometrika 32*;363—379.

——and H. F. KAISER(1964) Oblique factor analytic solutions by orthogonal transformations. *Psychometrika 29*;347—362.

HENDRICKSON, A. E. and P. O. WHITE(1964) Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology 17*;65—70.

HORN, J. L. (1965) An empirical comparison of various methods for estimating common factor scores. *Educational and Psychological Measurement 25*;313—322.

HORST, P. (1965) *Factor Analysis of Data Matrices*. New York: Holt Rinehart and Winston.

HOTELLING, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Education Psychology 24*;417—441, 498—520.

HOWE, W. G. (1955) Some Contributions to Factor Analysis. Report No. ORNL-1919. Oak Ridge, Tenn.: Oak Ridge National Laboratory. Ph. D. dissertation, University of North Carolina.

JENNRICH, R. I. (1970) Orthogonal Rotation Algorithms. *Psychometrika 35*;229—235.

——(1974) Simplified formulae in standard errors in maximum likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology 27*;122—131.

JENNRICH, R. I. and P. F. SAMPSON(1966) Rotation for simple loadings. *Psychometrika 31*;313—323.

JÖRESKOG, K. G. (1963) Statistical Estimation in Factor Analysis: A New Technique and Its Foundation. Stockholm: Almquist and Wiksell.

——(1966) Testing a simple structure hypothesis in factor analysis. *Psychometrika 31*:165—178.

——(1967) Some contributions to maximum likelihood factor analysis. *Psychometrika 32*: 443—482.

——(1969) A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika 34*:183—202.

——(1970) A general method for analysis of covariance structure. *Biometrika 57*:239—251.

——(1976) Analyzing Psychological Data by Structural Analysis of Covariance Matrices. Research Report 76—79. University of Uppsala, Statistics Department.

JÖRESKOG, K. G. and D. N. LAWLEY(1968) New methods in maximum likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology 21*:85—96.

KAISER, H. F. (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*:187—200.

——(1963) Image analysis, pp. 156—166 in C. W. Harris(ed.) *Problems in Measuring Change*. Madison: University of Wisconsin Press.

——(1970) A second-generation Little Jiffy. *Psychometrika 35*:401—415.

——(1974) Little Jiffy, Mark IV. *Educational and Psychological Measurement 34*:111—117.

——(1974) An index of factorial simplicity. *Psychometrika 39*:31—36.

KAISER, H. F. and J. CAFFREY(1965) Alpha factor analysis. *Psychometrika 30*:1—14.

KIM, J. O. (1975) Multivariate analysis of ordinal variables. *American Journal of Sociology 81*: 261—298.

——and C. W. MUELLER(1976) Standardized and unstandardized coefficients in causal analysis: An expository note. *Sociological Methods and Research 4*:423—438.

KIM, J. O., N. NIE and S. VERBA(1977) A note on factor analyzing dichotomous variables: the case of political participation. *Political Methodology 4*:39—62.

KIRK, D. B. (1973) On the numerical approximation of the bivariate normal(tetrachoric) correlation coefficient. *Psychometrika 38*:259—268.

LISREL III: Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods. (User's Guide). Jöreskog, K. G. and Sörbom, D. Chicago: National Educational Resources, Inc., 1976.

LITTLE JIFFY, MARK IV. (See Kaiser, 1974)

LABOVITZ, S. (1967) Some observations on measurement and statistics. *Social Forces 46*: 151—160.

——(1970) The assignment of numbers to rank order categories. *American Sociological Review 35*:515—524.

LAND, K. O. (1969) Principles of path analysis, pp. 3—37 in E. F. Borgatta(ed.) Sociological Methodology. San Francisco: Jossey-Bass.

LAWLEY, D. N. (1940) The estimation of factor loading by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh 60*:64—82.

——and MAXWELL, A. E. (1971) *Factor Analysis as a Statistical Method*. London: Butterworth and Co.

LEVINE, M. S. (1977) Canonical Analysis and Factor Comparison. Sage University Papers on Quantitative Applications in the Social Sciences, 07—006. Beverly Hills and London: Sage Pub.

LI, C. C. (1975) *Path Analysis—A Primer*. Pacific Grove, Calif.: Boxwood Press.

LINN, R. L. (1968) A Monte Carlo approach to the number of factors problems. *Psychometrika 33*:37—71.

LORD, F. M. and W. R. NOVICK(1968) *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

MALINVAND, E. (1970) *Statistical Methods of Econometrics*. New York: Elsevier.

MAXWELL, A. E. (1972) Thomson's sampling theory recalled. *British Journal of Mathematical and Statistical Psychology 25*:1—21.

McDONALD, R. P. (1970) The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology 23*:1—21.

——(1974) The measurement of factor indeterminacy. *Psychometrika 39*:203—221.

——(1975) Descriptive axioms for common factor theory, image theory and component theory. *Psychometrika 40*:137—152.

——(1975) A note on Rippe's test of significance in common factor analysis. *Psychometrika 40*: 117—119.

——and E. J. BURR(1967) A comparison of four methods of constructing factor scores. *Psychometrika 32*:380—401.

MULAIK, S. A. (1972) *The Foundations of Factor Analysis*. New York: McGraw-Hill.

NEUHAUS, J. O. and C. WRIGLEY(1954) The method: an analytic approach to orthogonal simple structure. *British Journal of Mathematical and Statistical Psychology 7*:81—91.

OSIRIS Manual. Ann Arbor, Mich.: Inter-University Consortium for Political Research, 1973.

RAO, C. R. (1955) Estimation and test of significance in factor analysis. *Psychometrika 20*: 93—111.

RUMMEL, R. J. (1967) Understanding factor analysis. *Conflict Resolution 11*:444—480.

——(1970) *Applied Factor Analysis*. Evanston: Northwestern University Press.

SAS: A User's Guide to SAS 76. Anthony J. Barr, James H. Goodnight, John P. Sall, and Jane T. Helwig. Raleigh, N. C.: SAS Institute, Inc., 1976.

SPSS: *Statistical Package for the Social Sciences*. Norman H. Nie, C. Hadlai Hull, Jean G. Jenkins, Karin Steinbrenner, and Dale Bent. New York: McGraw-Hill, 1975.

SAUNDERS, D. R. (1953) An Analytic Method for Rotation to Orthogonal Simple Structure. Research Bulletin 53—10. Princeton, N. J.: Educational Testing Service.

——(1960) A computer program to find the best-fitting orthogonal factors for a given hypothesis. *Psychometrika 25*:199—205.

SCHUESSLER, K. (1971) Analyzing Social Data. Boston: Houghton Mifflin.

SÖRBOM, D. and K. G. JÖRESKOG(1976) COFAMM: Confirmatory Factor Analysis with Model Modification User's Guide. Chicago: National Educational Resources, Inc.

STEPHENSON, W. (1953) *The Study of Behavior*. Chicago: The University of Chicago Press.

STEVENS, S. S. (1946) On the theory of scales of measurement. *Science 103*:677—680.

STINCHCOMBE, A. L. (1971) A heuristic procedure for interpreting factor analysis. *American Sociological Review 36*:1080—1084.

THOMPSON, G. H. (1934) Hotelling's method modified to give Spearman's g. *Journal of Educational Psychology 25*:366—374.

THURSTONE, L. L. (1947) *Multiple Factor Analysis*. Chicago: University of Chicago Press.

TRYON, C. R. and BAILEY, D. E. (1970) *Cluster Analysis*. New York: McGraw-Hill.

TUCKER, L. R. (1966) Some mathematical notes on three mode factor analysis. *Psychometrika 31*:279—311.

——(1971) Relations of factor score estimates to their use. *Psychometrika 36*:427—436.

——R. F. KOOPMAN, and R. L. LINN(1969) Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika 34*:421—459.

TUCKER, L. R. and C. LEWIS(1973) A reliability coefficient for maximum likelihood factor analysis. *Psychometrika 38*:1—8.

VELICER, W. F. (1975) The relation between factor scores, image scores, and principal component scores. *Educational and Psychological Measurement 36*:149—159.

WAINER, H. (1976) Estimating coefficients in linear models: it don't make no nevermind. *Psychological Bulletin 83*:213—217.

WANG, M. W. and J. C. STANLEY(1970) Differential weighing: a review of methods and empirical studies. *Review of Educational Research 40*:663—705.

译名对照表

a graphical rotation	图像旋转
a principal axis factoring	主轴因子分析
a quartimin criterion	四次方最小标准
a scale-free extraction method	无尺度的抽取方法
a target matrix	目标矩阵
a trivariate normal distribution	三元正态分布
Alpha factoring	Alpha 因子分析法
an observed covariance structure	观察到的协方差结构
anti-image	反映像
biquartimax	二分四次方最大法
biquartimin	二分四次方最小
bivariate normal	二元正总分布
canonical correlation	正则相关系数
canonical factoring	正则因子分析法
characteristic equation	特性方程
characteristic roots	特征根
common factor model	共同因子模型
common factors	共同因子
component scales	成分尺度
component scores	成分得分
covarimin	最小协方差
determinantal equation	行列式方程
direct oblimin	直接最小斜交法
eigenequation	特征方程
eigenvalues	特征值
empirical confirmation	经验证实
empirical constraints	经验限制
equimax	均等变化法
error variance	误差方差
extrastatistical	与统计无关的
factor loadings	因子负载
factor pattern matrix	因子模式矩阵
factor scales	因子尺度
factorial causation	因子因果关系
factorial complexity	因子复杂性

generalizability coefficient	概化系数
hyperplanes	超平面
image analysis	映像分析
image factoring	映像因子分析法
image factors	映像因子
image	映像
indicator variables	指示性变量
informativeness	资讯性
invariance	恒定性
kurtosis	峰度
Likert scale	李克特量表
linear system	线性系统
maximum fit	最大拟合
metric variables	度量变量
minimum discrepancy	最小差异
minres	最小化残差法
Monte Carlo experiments	蒙特卡洛实验
nonmetric	非度量的
oblique rotation	斜交旋转
orthoblique method	正斜交旋转法
parsimony	简洁
partial anti-images	偏反映像
partial images	偏映像
principal axis factoring with iterated communalities	带迭代公因子的主轴因子分析法
principal components analysis	主成分分析
promax oblique rotation	最优斜交旋转法
psychometric sampling	心理测量抽样
quartimax	四次方最大法
reference axes	参照轴
reliability	信度
response sets	响应组
sampling variability	抽样变异性
scale scores	尺度得分
scree-test	碎石检验
simple factor structure	简单因子结构
subspaces	子平面

tetrachoric correlations	四分相关系数
the degree offactorial determination	因子的确定程度
the primary pattern matrix	主模式矩阵
the rank-theorem	秩理论
the residual-mean-square	残差均方
the total image	总映像
the underlying causal structure	潜在的因果结构
the underlying factor pattern	潜在的因子模式
unique components	独有成分
unique factors	独有的因子
unique variance	独有方差
univocal	单一的
varimax	最大方差法
vector spaces	向量空间

作 者 简 介

奈杰尔·吉尔伯特(Nigel Gilbert)

英国吉尔福德萨里大学社会学教授,《虚拟社会期刊》(*Journal of Artificial Societies*)和《社会模拟》(*Social Simulation*)的编辑。研究领域主要是基于行动者模型对理解社会和经济现象的应用,特别是规范、文化和创新的产生。社会模拟领域的先驱之一,也是欧洲社会模拟学会的前会长,英国社会科学院院士和皇家工程院院士。

译 者 简 介

盛智明

曾于2008年和2010年分别获北京大学社会学硕士学位和香港科技大学社会科学部硕士学位,目前是香港科技大学社会科学部的博士研究生,研究方向主要是当代中国的社会运动、政治参与和基层政治改革。

序

研究社会行为有两种普遍的方法:收集观察抽样调查的或其他形式的数据,并通过估计一个模型来分析它们;或者是从关于某种社会行为的理论认识开始,建立一个关于它的模型,然后模拟它的动态过程以获得一个对看似简单的社会系统的复杂性的更好理解。Sage“社会科学中的定量应用”系列丛书中的大部分采用的都是第一种方法。基于行动者的模型,作为极少的例外之一,采用的是第二种方法。

应用于社会科学的许多统计方法都有一个漫长的发展历史。比如,回归方法起始于高尔顿(Galton)在19世纪的开创性工作。基于行动者模型的起源通常归功于冯·诺依曼(Von Neumann),因为他在20世纪40年代晚期发明了以他的名字命名的机器。虽然名为冯·诺依曼的机器并不是基于行动者模型甚至也不是任何种类的模型,但是它们作为理论建构为虚拟生活的建构和模型化奠定了基础。在过去的十几年里,基于行动者模型在经济学、政治学、社会学和其他一些社会科学中已经变得越来越流行。

这一分析方法有何独特之处呢? 比如,以Thomas Schelling发表在1971年《数学社会学期刊》(*Journal of Mathematical Sociology*)的经典之作为例。他建立了一个非常简单的区隔动态模型,这个模型是通过在棋盘上依照一定的简单规则去移动一分和一角的硬币来实现的。一分和一角的硬币代表了居住在棋盘(城市)方格(房屋或地段)中两个来自不同社会群体的行动者。令人惊奇的发现在于,虽然每个行动者都容忍来自不同群体的邻居(或者适度地偏好自己所属的群体),但总体(在整个棋盘—城市中)最终依然以出现严重的区隔结束。Schelling这个简单的基于行动者模型揭示出了微观层次的个体动机和宏观层次的社会行动之间的一些复杂关系。

Nigel Gilbert,作为一名基于行动者模型的国际权威和《虚拟社会和社会模拟期刊》(*Journal of Artificial Societies and Social Simulation*)的编辑,是为这个系列的此主题写书的最合适人选。在本书中,Gilbert以通篇的例子来解释基于行动者的模型的基本原理并讨论了基于行动者模型的应用软件。他写的清晰、明确,这对于使一个不熟悉的议题能够被研究者掌握是非常重要的,这些研究者都能熟练地运用回归类型方法。这本书及时地填补了这个系列的重要空白,这个空白是我作为系列丛书的编辑希望填补的,它也有助于传播一种在本质上非常有用的方法,而这种方法是我们作为社会科学家都应该知道的。

——廖福挺

前 言

基于行动者的模型是一种计算机模拟的形式。虽然模拟作为一种研究技术在从天文学到生物化学的自然科学中已经是非常重要的部分，但直到最近为止它在社会科学中仍然是被忽视的。这可能是因为缺少尊重社会科学特殊需求的计算机方法。然而，在 20 世纪 90 年代早期，基于行动者的模型的价值开始逐渐被认识，从那以后，运用基于行动者的模型的研究数量开始迅速增长，一些例子也开始在核心期刊上出现(Gilbert & Abbott, 2005)。

当理解一些议题的过程和它们的结果同样重要的时候，基于行动者模型就特别适合于这些议题。在本质上，人们创建一个计算机程序，然后运行这个程序，观察它在整个模拟期间在进行什么。在被模型化的行动者和程序中的行动者之间存在一个直接的对应，这使得这种方法在直观上颇具吸引力，尤其是对那些玩电脑游戏长大的一代。尽管如此，基于行动者的模型从严谨性上来说仍然位于数学和统计模型一边。如同基于公式的模型，而不是像散文那样。基于行动者模型如果能够被电脑执行的话，它们就必须具有完整性、一致性和清晰性。另一方面，基于行动者的模型和大多数数学模型不同，它能够包含在特征和能力上有异质性的行动者，能够对远离均衡的情况建模，并能够直接处理行动者之间互动的结果。

因为是一种新的方法，所以现在只有很少的课程讲授基于行动者的模型的建模技术，虽然课程数量正在迅速增加，但是只有很少的课本是直接针对有兴趣的社会科学研究者的(Degushi, 2004; Gilbert & Terna, 2000; Gilbert & Troitzsch, 2005; Grimm & Railsback, 2005; Liebrand, Nowak & Hegselman, 1998; Macy & Willer, 2002; Schweitzer, 2003; Tesfatsion & Judd, 2006)。

这本书提供了对此模型的介绍，强调了当选择基于行动者模型作为合适的方法时建模者需要作出的决定，并给出了关于如何进行相关操作的一些技巧。本书的目的是使社会科学家和研究生们能够进行操作实践，而且它已经在计算机社会科学研究生层次的模块或博士生项目中被用做基于行动者建模的推荐读物。它也适合于作为关于高级社会研究方法的研究生课程中的辅助性读物。在任何语言下的计算机编程知识和经验都将有助于理解本书，但并不是必要的。

这本书最后以出版的书籍、网络资源和参考文献作为结束。因为这个领域发展得异常迅猛，所以只能提及现有研究中的一些例子和在特定议题上提供有着更多细节的参考书。如果容量足够的话，其实还可以引用更多的内容。另外，本书只简要地提及了两个与基于行动者的模型紧密相关的领域：网络模型和博弈模型，关于二者更加详细的内容都在 Sage 丛书的其他卷中有涉及，比如 Scott 的社会网络分析(*Social Network Analysis*)(2000)以及 Fink，Gates 和 Humes 的博弈理论问题(*Game Theory Topics*)(1998)。

第 1 章 | 基于行动者建模的构想

基于行动者的模型虽然对于社会科学而言是一个新的分析方法，但却迅速地流行起来。这本书解释了基于行动者的模型到底是什么，它也提醒了在运用此模型时存在的一些危险，并描述进行基于行动者模型建模的一些典型方法。最后，它提供了一系列来自社会科学的例子。

第 1 章在将基于行动者的模型与其他可能更令人熟悉的模型对比之前，先简要概述了此模型，并描述了目前在基于行动者建模研究中的一些例子。第 2 章将介绍更加细节的部分，思考一系列方法论和理论问题，并解释什么是"行动者"。第 3 章进入到建立基于行动者模型的具体环节当中，回顾可用的软件平台，并逐步展示如何使用其中之一来建立基于行动者的模型。第 4 章提供了关于如何设计基于行动者的模型，如何在社会科学研究中使用它们以及如何发表以基于行动者建模为基础写成的论文的一些切合实际的建议。第 5 章讨论了基于行动者建模研究的未来发展和可能取得的进展会在哪里。最后，本书以相关的网络和出版资源列表作为结束。

基于行动者建模作为社会科学中的一种建模方法已经变得越来越流行，因为它能够建立代表单个实体和他们之间互动的模型。在与运用结构方程的基于变量的方法，或与运用微分方程的基于系统的方法进行比较时，基于行动者建模为模型化个体的异质性提供了可能，它能明确地表示出行动者的决策规则，并将行动者放置于地理或其他类型的空间中。它允许建模者以一种自然的方式表示出分析的多种尺度，从个人行动到宏观或社会层次结构的出现，以及适应和学习的各种种类，这些都是其他建模方法所不能轻易做到的。

基于行动者建模

基于行动者建模是一种计算机方法，它能够使研究者创建，分析和试验由在环境中互动的行动者构成的模型。让我们分析一下在这个定义中的每一个关键词。

一种计算机方法

首先，基于行动者的模型是计算机社会科学的一种。也就是说，它涉及建立作为计算机程序的模型。建模的思想在大多数社会科学中都为人所熟知：创建某种“社会现实”的简化代表，以尽可能清晰地展现出人们所相信的现实世界的运行方式。例如，如果有一个因变量和一个或更多的自变量，回归方程就可以作为变量之间关系的模型。由节点和边线组成的网络可以模型化一组朋友关系。甚至对一种关系的普通语言描述，比如一个国家中保护知识产权的力度与创新程度之间的关系，也可以被看做是一个模型，虽然这是一个简单和相当非正式化的模型。

计算机模型作为计算机程序被公式化，其中存在一些输入（如自变量）和一些输出（如因变量）。程序本身代表了被认为存在于社会世界中的一些过程（Macy &Willer, 2002）。例如，我们可能有一种关于朋友影响消费者购物选择的理论。我们将会看到，我们能建立一个程序，在此程序中，个人（行动者）根据他们自己的偏好来购物。而有趣的地方在于，一个行动者所购买的物品将影响其朋友的购物行为，并且这个朋友所购买的物品又会反过来影响到第一个行动者。运用基于行动者建模就相对比较容易模型化这种相互强化的过程。

计算机建模的一个好处是它迫使人们要精确：不像用自然语言表达的理论和模型，如果要使一个计算机程序运行的话，那这个程序就必须是完整的和完全具体化的。另一个优点在于一般而言它相对比较容易能模型化关于过程的理论。如果建构计算机模型的构想使你想起了电脑游戏，特别是玩家自己去构

建虚拟世界的那种，例如 Sims(http://thesims.ea.com)，这也并不是巧合。这些游戏其实非常接近计算机建模，虽然为了使它们变得有趣，这些游戏通常会比基于行动者的模型具备更新奇的画面和更少的社会理论。

试验

在物理学、化学以及部分生物学中，试验是进行科学研究的标准方法，但在社会科学中，进行试验是不可能或不合适的。一个试验包括将一些处理应用于一个封闭的系统并观察发生的反应，将被处理过的系统与另一个在其他方面都相同但未受到处理(被控制)的系统做比较。试验的最大优点在于使人们能够确认正是这个处理引发了被观察到的效应，因为在被处理系统和被控制系统之间不同的只有这个处理，并且这些系统是与其他可能引发变化的因素隔离开来的。然而，在社会系统中，隔离通常是不可能的。因此，虽然试验的结果具有潜在的清晰性，但却很少被社会科学家所采用就不奇怪了。

基于行动者建模的一个最大优点是：在确保人类系统的隔离和避免试验的道德困境时可能会遇到的困难在对虚拟或计算机系统进行试验时都将不存在。试验可以通过使用一系列参数或允许一些因素随机变化的方法而被配置和重复很多次。当然，用关于一些社会现象的计算机模型来进行试验时，只有当这个模型以与人类系统一样的方式运作，或换句话说，只有当这个模型是一个好模型的时候，才会产生有趣的结果，但人们可能并不知道情况是否如此。所以，基于模型的试验并非是万能的。

在模型中而不是在真实系统中进行试验的想法其实并不新颖。例如，建筑师将一个摩天大楼的模型放在风洞中去测试它在狂风下的运动，他们对模型进行试验的原因与社会科学家想要对他们的模型进行试验的原因是一样的：对真实摩天大楼进行试验的成本太高了。另一个对模型进行试验的理由在于这可能是得到结果的唯一方法。分析性地得出模型的运动通常是最好的方法，因为它在一系列给定的输入值下给出了关于模型将如何运动的信息，但通常分析性的解决方案并不容易获得。在这些情况下，就有必要对不同的输入值进行试验来看模型是如何运动的。模型可能处于许多不同的环境之

下，以此来模拟真实世界。

模型

计算机社会科学是基于建构模型然后用它们来理解社会世界这样的思想之上的(Sawyer，2004)。模型在社会科学中已有很长的历史——比计算机使用的历史还要长很多——但只有在统计方法开始应用于分析经济学和人口学中的大量定量数据时它才崭露头角。模型试图去代表或模拟一些真实存在的现象，这被称为模型原型。模型的两个主要优点在于它简明地展现了原型特征之间的关系，并且它使人们可以通过研究模型发现关于原型的信息(Carley，1999)。

一个众所周知的最早的社会科学模型是 Phillips(1950)的经济学水力模型，在这个模型中，相互联系的管道中流动的水被用来代表货币的流通。这个模型目前仍然能够在伦敦的科学博物馆中被欣赏到(http://en.wikipedia.org/wiki/MONIAC_Computer)。改变参数诸如利息率的影响可以通过改变流经管道的水的流动率来研究。

模型有许多风格，有必要列出其中的一些来说明它们之间的差异：

(1) 比例模型是原型较小的版本。与大小一起减小的是在模型细节和复杂性水平上的系统性减小。例如，一架飞机的比例模型可能与它的原型具有相同的形状，但可能没有展现电子控制系统或可能没有真飞机上的引擎。同样，一个城市的比例模型将比真实的城市小很多而且可能只模型化了两个维度(例如，建筑物之间的距离，却没有建筑物的高度)。当在研究完模型后再对原型下结论时，我们需要记住，从模型中得出的结论需要按比例还原到原型的维度上去，并且有些没有被模型化的特征很可能会影响到结论的有效性。

(2) 理想型模型是原型中的一些特点被夸大以便于简化的模型。例如，一个股票市场的理想化模型可能假设信息能够及时地从一个交易者流向另一个交易者；交通的理想化模型可能假设司机从来不迷路等等。理想化具有将一个或更多复杂因素移除出模型的效应，如果这对模型如何运行的影响微乎其微，那么这个模型得出的关于原型的结论仍将是有用的。

(3) 类比模型是以一些已经得到很好理解的现象与原型进行类比为基础的模型。最著名的例子就是原子的撞球模型，但社会科学中也存在这样的例子，例如思维的计算机模型(Boden, 1988)和组织的垃圾箱模型(Cohen, March & Olsen, 1972)。这些模型都是有用的，因为从类比中得到确认的结果能够被延续和应用到原型上，当然这些结果的有效性依赖于类比的恰当性。

以上这些类别都不是互斥的，一个可能的并且的确很普遍的情况是一个模型可能同时是比例模型和类比模型(例如，上面提到的经济学水力模型就是这样的混合)。

有些模型归于与前面都有些不同但在社会科学中也会经常遇到的第四类别，这些模型通常被称为数学或基于公式模型。例子就是定量社会学中的结构方程模型和新古典经济学中的宏观经济模型。这些模型具体化了变量之间的关系，但与其他三个类别的模型不同，它们并没有显示模型和原型之间的类比或相似。通常，数学模型的成功是通过数据与公式的拟合程度表现出来的，而公式形式本身却不令人感兴趣或重要性不大。例如，Cobb-Douglas 的生产函数是一个关于生产产出如何与生产投入相关的数学模型(Cobb & Douglas, 1928)：

$$Y = AL^{\alpha}K^{\beta}$$

其中 $Y =$ 产出，$L =$ 劳动力投入，$K =$ 资本投入，A、α、β 是由技术决定的常数项。这个公式的形式来自于统计证据，而不是来自于对关于公司行为的理论化。虽然数学模型在阐明变量间关系的社会科学的某些方面非常成功，但它们在帮助人们理解为什么一个变量会与另一个变量相关这点上通常不是很有用，或者换句话说，在表达关于过程和机制的构想上不是很有用，而在这点上其他种类的模型通常会更有帮助。

行动者

基于行动者的模型由在一个环境中互动的行动者构成。行动者或者是单独的计算机程序，或者是更普遍的用来代表行动者的一个程序的不同部分——

单个的人，组织如公司，或实体如民族—国家。它们被编入程序中来对它们所处的计算机环境作出回应，而这个环境是社会行动者在其中行动的真实环境的模型。

正如在后面我们将会看到的，基于行动者模型的一个关键特征在于行动者能够进行互动。也就是说，他们能相互传递信息性消息，并且能够基于他们从这些消息中学习到的内容展开行动。这些消息可能代表人们之间的口头对话或信息流动的更加非直接的方式，例如对另一个行动者的观察或对另一个行动者行动影响的检测。这种模型化行动者与行动者互动的可能性是将基于行动者模型与其他种类的计算机模型区分开来的主要方式。

环境

环境是行动者在其中行动的虚拟世界。它可能是一个完全中立的媒介，对行动者产生很小的影响或没有影响，或者在其他模型中，环境可能如行动者自身一样被小心地制作出来。通常，环境代表地理空间，例如，在关心居住隔离的模型中，环境模拟出一个城市的一些物理特征。在国际关系的模型中，环境绘制出国家和民族（Cederman，1997）的版图。那些环境在其中代表地理空间的模型被称做空间直观。在其他模型中，环境可以是一个空间，但并不表示地理环境而是其他一些特征。例如，科学家可以被模型化进“知识空间”（Gilbert，Pyka & Ahrweiler，2001）。在这些空间模型中，行动者用坐标来表明他们的位置。另一种选择是没有空间表征但是将行动者全部连接进一个网络中，在这个网络中一个行动者与其他行动者之间关系的唯一标识就是用网络连线与之连接起来的行动者列表（Scott，2000）。

为了使这些定义在某种程度上变得更加具体，在下一节中，我们将介绍关于这些概念的基于行动者模型的一些例子。

一些例子

基于行动者的模型在社会科学的大多数分支学科中都具有价值。这一节

所简要介绍的模型是选择出来说明它们涉及的问题领域的多样性。在这些领域中它们被卓有成效地加以应用。

城市模型

1971年，Thomas Schelling(1971, 1978)，也可参见Sakoda(1971)提出了一个模型，用于解释在美国城市中观察到的种族隔离。这个模型作为原创性构想是一个非常抽象的模型，但它在解释美国和其他城市中心的种族隔离持续性的新近研究中已经很有影响力。这个模型基于一个常规的由单元格组成的正方形网格，正方形网格代表一个城市区域，在此区域内的行动者代表家庭户，他们是随机安置的。行动者有两种(让我们称他们为"红"和"绿")。每个单元格一次只能容纳一户家庭行动者，并且有许多单元格是空的。在每一个时间步骤中，就每户家庭对他直接的邻居(围绕着他的8个单元格)进行调查并算出与他不同颜色家庭的分数。如果这个分数比一些恒定的"容忍"阈值大，即比围绕在一个绿格周围的已设定的红格比例或在一个红格周围的已设定的绿格比例大，这个家庭就会认为自己"不幸福"并决定搬迁。他就会迁移到网格中一些空白的单元格里。

在下一个时间步骤中，这个新安置的家庭可能会打破他邻居的"容忍"平衡，使他们中的一些变得"不幸福"，这就可能导致一连串的重新安置。当容忍阈值在0.3或以上时，最初随机分布的家庭户就分隔成红色和绿色的小块，每种相同颜色的家庭聚集到一起(图1.1)。即使当家庭可以"容忍"居住在临近的大

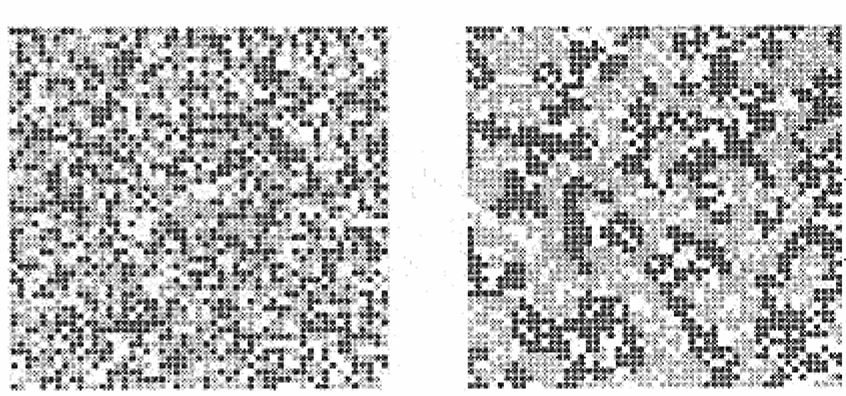


图1.1 开始时(左)和达到均衡后(右)的Schelling模型，0.3的统一容忍度

资料来源：Wilensky, U. (1998), NetLogo隔离模型，http://ccl.northwestern.edu/netlogo/models/segeration.关联学习和基于计算机模型中心，西北大学，Eranston, IL。

部分邻居都是不同颜色时，这种聚集也会发生。Schelling 解释说，这表明了即使种族偏见的程度很低，也有可能产生像 20 世纪 70 年代美国城市那样的典型的强隔离模式。

Schelling 模型之所以具有影响力是因为以下几个原因（Allen，1997）。首先，其结果——相同颜色的家庭聚集——令人惊讶，并且这一结果如果只从个体行动者的运动规律考虑的话是不容易预料到的。第二，模型非常简单，只有一个参数值即"容忍"阈值。因此是很容易理解的。第三，意外的聚集行为是相当稳健的。在许多不同的容忍阈值下，在不同的运动规律下（例如，家庭行动者能够随机选择一个新单元格，或者使用效用函数选择最喜欢的单元格，或者如果单元格标明了价格，则将支付能力也纳入考虑范围，等等），以及在关于应考虑哪些邻居的不同规则下（例如，在周围 8 个单元格中；东南西北 4 个单元格中；或远处更大的两圈或更多的单元格）（Gilbert，2002），都能获得相同的结果。第四，这个模型直接给出了如何用经验数据检验它的建议（Benenson，Omer & Hatna，2002；Bruch & Mare，2006；Clark，1991；Pancs & Vriend，2004；Zhang，2004），虽然实践证明非常难取得关于家庭地点偏好的可靠和广泛的数据来计算"不幸福"比率。Schelling 模型与前面提到的其他基于与迁移流动和居住物业相对价值有关的公式的模型相比，其优点在于需要估计的参数数目更少并且模型易于模拟和分析。现在工作就集中于将此模型变得更加具体，将抽象的正方形网格替换为真实的城市地理，并增加更多的因素，例如家庭对想要迁入地点的支付能力。

民意动态

另一组具有重要潜在意义的有趣模型组是关于对政治民意发展的理解。例如：解释在一个群体中极端主义民意的扩散。虽然有许多相关的研究探讨了不同假设的结果和民意传播的机制，但我们在此只回顾一个这样的研究（Deffuant，2006；Deffuant，Amblard & Weisbuch，2002；Hegselmann & Krause，2002；Lorenz，2006；McKeown & Sheehy，2006；Stauffer，Sousa & Schulze，2004）。Deffuant 等（2002）问到：

> 最初被认为是极端和边缘的民意如何能成功地变为一个群体中大部分人的准则？人类历史上的一些例子表明，一个大型社区由于在刚开始时受一小撮少数人的影响会或多或少有点突然地在全球范围内形成一种极端民意。20世纪30年代的德国就是关于这个过程的一个特别富有戏剧性的例子。在最近的几十年中，一群最初为少数派的极端伊斯兰分子成功地说服了中东国家的大部分人口。我们可以考虑一些不那么具有戏剧化的过程，比如时尚，这就是少数人群体的行为，一旦其被认为是前卫分子，就会成为大部分人的准则。另一方面，人们也可以找到许多产生了严重的群体两极分化的例子，比如冷战时期的欧洲。在这些例子中，整个人群都变成了极端派（选择一边或另一边）。

在Deffuant等的模型中，行动者持有一种观点（一个在−1.0和+1.0之间的真实数字）以及对他们自己观点的怀疑度，称做不确定性（一个正的真值）。行动者的观点分段被界定为行动者观点中心的一段，随着行动者不确定性值的大小向右或向左延伸。行动者随机互动。当他们相遇时，如果他们的观点分段重叠，一个行动者可能就会影响另一个行动者。如果观点分段不重叠，就假设行动者的观点差异较大以至于他们没有机会影响彼此。如果一个行动者影响了另一个行动者，那么这个行动者（行动者j）的观点受另一个行动者（行动者i）观点影响的大小就是他们之间观点差异的一定比例数，乘以观点重叠数除以行动者i的不确定性程度，再减1。这个公式的效应是，具有非常不确定性的行动者对其他行动者产生的影响要比那些具有确定性的行动者对其他行动者产生的影响小（Deffuant，et al.，2002，公式1到6）。

除了一些拥有最积极和最消极观点的极端派，每一个行动者在刚开始时都持有一个来自于统一随机分布的观点，并具有相同层次的不确定性。极端派被给予一个低水平的不确定性，即假设极端派对他们的极端观点相当确定。在这些条件下，极端主义扩散开来，并且最终模拟达到了一个稳定状态，所有行动者都成为了在民意连续体的一端或另一端的极端分子。在不存在具有政治确定性的极端分子的情况下重新进行模拟，取而代之的是群体融合，所有行动者共

享一个温和的观点。因此，模型表明一些其观点不能被其他行动者所影响的极端分子能够对大多数人的观点产生戏剧性的影响。这一研究对于理解恐怖主义运动的发展有一定的启示，在这些运动中一些极端分子能够从更广泛群体的相当比例的人口中获得支持。

消费者行为

对商业界而言，自然非常想知道究竟什么会影响消费者购买他们的产品。虽然产品内在的质量通常是重要的，但朋友、家庭、广告、时尚和其他一系列“社会”因素的影响也同样重要。为了考察这些因素之间复杂的互动，一些研究者已经开始运用基于行动者的模型，其中行动者代表消费者。在这些研究中，第一个报告这一模型的是 Janssen 和 Jager(1999)，他们探讨了导致“锁定”消费者市场中的过程。当在许多竞争性产品中的某一个实现了主导以至于对于个体消费者而言很难转向其他竞争者的产品时，锁定就发生了。经常被引用的例子是 VHS 录像带(超越了 Betamax)，QWERTY 键盘(超越了其他按键布局)，微软操作系统(超越了 Apple 和 Linux)，等等。Janssen 和 Jager 关注于导致锁定的行为过程，因此他们为被他们称为“消费者”的行动者赋予了心理上可能的决策规则，并仔细地依据社会比较和模仿等行为理论对其进行合理化。

另一个模型化消费者的例子是 Izquierdo 和 Izquierdo(2006)所做的研究，在这个研究中，作者考虑的市场，如二手车市场，具有质量可变性(不同的产品有不同的质量)和质量不确定性(在购买和使用一个产品之前很难知道它的质量)。此研究探讨了质量可变性是如何损害市场和影响消费者信心的。这里有两个行动者角色：买方和卖方。卖方通过计算他们可以接受的最低价格出卖商品，买方基于商品的期望质量给出一个价格购买商品。期望质量是基于经验，要么来自于行动者自己的经验，要么来自于对社会网络中行动者同辈群体经验的积累。系统中的商品数量是有限的，买方和卖方在每一轮买卖中完成一笔交易之后市场都要被清空。社会网络是通过随机连接一对行动者建立起来的，用一个参数来调整连接的数量，从完全连接到完全分开。

作者发现，如果没有社会网络，消费者信心就会降到市场无法再维持的临

界点之下，但如果有社会网络，那么行动者自身经验的积累和其他行动者更多积极集体经验（不是反复无常的）的积累就有助于维持市场的稳定。这表明了社会信息如何能使群体经验积累到一个更准确的水平从而降低单个个体不良经验的重要性。

产业网络

大多数经济理论忽略了公司间联系的重要性，但有很多产业区的例子可以说明网络很重要。一个众所周知的例子当属意大利北部的"工业区"，如纺织品生产区普拉托(Prato)。工业区的特点是大量的小公司在一个较小区域内聚集在一起，它们都生产相同种类的产品，相互之间有很强但又有差异的联系。这些联系可能是供应商/消费者的联系，共享技术的合作关系，财务关系，或只是朋友或亲属关系(Albino, Carbonara & Giannoccaro, 2003; Boero, Castellani & Squazzoni, 2004; Borrelli, Ponsiglione, Iandoli & Zollo, 2005; Brenner, 2001; Fioretti, 2001; Squazzoni & Boero, 2002)。另一个例子是"创新网络"，这在知识密集型部门，如生物技术和信息技术部门比较普遍。属于这些经济部门的公司并不总是在地理上聚集在一起（虽然在某些区域有集中的趋势），但它们的确与其他类似的公司有很强的网络联系，与这些公司共享知识、技能和技术。

例如，Gilbert 等提出了一个创新网络模型，其中行动者通过"kenes"来象征他们的知识和专业股份，kenes 被用来发展新产品以在市场上向模型中的其他公司销售。然而，一个产品只有在它的组件可以从其他公司那里购买到，并且要有某些公司想购买这个产品时，它才能被生产出来。因此，在某种程度上，这个模型是一个由相互进行贸易的公司组成的产业市场的模型。另外，公司可以通过内部研发或者通过合作安排从其他公司吸纳获取知识的方式来提高他们的 kenes。改进的知识可以用来生产能够卖得更好的产品，或只需要更少或更便宜的组件。在某种程度上，这个模型与能够通过一种自然选择（参见第 5 章第 2 节）而学习的群体相类似，在这个自然选择中，那些不能找到消费者的公司将停止交易，而最优秀的公司将生存下来，与其他公司进行合作，并

产生把它们“母”公司最好的方面吸纳进去的子公司。另一个例子参见 Pajares，Hernández-Iglesias 和 López-Paredes(2004)。

供给链管理

制造商通常从其他组织那里购买组件并将他们的产品卖给分销商，分销商再卖给零售商，最终，产品到达使用者手中。而使用者可能并不知道协作递送产品的复杂组织间关系。当产品涉及更多组成部分，要在全世界范围内更广泛地进行调拨，以及当管理者试图减小库存、增加货物可得性时，最大化使商家连接起来的供给链的效率变得越来越重要，也越来越困难。模型化供给链是研究订单履行过程和考察管理政策有效性的一个很好的途径。多行动者模型被越来越多地运用于这一目的。

多行动者模型非常适合模拟供给链的任务，因为涉及的商业部门能够被模型化为行动者，每个行动者也都有自己的存货规则。同时也容易模型化产品沿着供给链向下流动和信息从一个组织向另一个组织的流动，比如订货量和领导时间。这就是 Strader，Lin 和 Shaw(1998)所采用的方法，他们描述了他们为研究在分散的装配供给链中被共享的信息的影响而建立起来的一个模型。分散的装配供给链在电子和电脑产业中较为典型，在这一链条中有一小部分数量的供应商提供材料和次级组件(如电子仪器)用来组装非商标产品(如硬盘驱动器)，然后再用这些产品在销售地组装成定制产品。Strader 等人比较了 3 种订单执行政策：(1)依据订单制造，生产由客户订单引发；(2)依据订单装配，组件已制造好并存放在库存中，只有最后的装配是由订单引发；(3)依据库存制造，生产是受库存量跌至一定的界限之下所引发。他们也对组织之间共享信息的不同数量进行实验，发现在他们模型化的分散组装供给链中，依据订单装配这一策略，与在供给链的组织之间共享供给和需求信息一起作用是最有效率的。他们也指出，其结果支持这样一种观点，即信息可以代替库存：如果人们掌握准确的信息，就能降低关于需求的不确定性，其结果就是，用于满足订单所需要的库存水平也能相应降低。

电力市场

在许多发达国家，最近一些年，电力供应开始私有化。通常是由 2 个或 3 个电力公司先将电力卖给一些销售商，这些销售商再将电力卖给商业或家庭用户。从一个垄断的、国有或国家控制的供应商到一些供应公司通过竞标进入市场，这样的转变已经启发一系列基于行动者的模型想要去预测市场调节的影响；供应商和购买者数量和种类的变化；以及预期的政策变化，比如减少电力中断的次数或减小电力生产对环境的影响(Bagnall & Smith，2005；Batten & Grozev，2006；Bunn & Oliveira，2001；Koesrindartoto，Sun & Tesfatsion，2005；North，2001)。

在这些模型中，行动者是电力供应公司，他们在一段时间内，如一天或一小时内，以一个确定的价格向模拟的市场出售一定数量的电力。现实中电力市场也是如此运行的：公司对电力供应报价，最好的报价被接受(不同的市场在关于什么是"最好"的报价上有不同的准则)。通常对电力的需求是连续变化的，所以对电力供应公司而言，设定电力价格以最大化他们的利润是一个困难的工作。一个更加复杂的情况是生产的成本可能是非线性的：匹配顶峰用电需求可能意味着要开动一个只需要使用几小时的电力站。

通过运行模拟，可以研究市场价格降到接近生产边际成本的条件；减少电力供应公司数量的兼并者的影响；以及不同种类的市场"设计"的结果，诸如允许期货交易。目前大多数的模型允许行动者通过使用一种被称之为"强化学习"的技术去"学习"交易策略(参见第 5 章第 2 节)。一个供应商行动者从进行投标开始，投标所用的价格策略就是从对所有供应商而言都通用的一组价格策略中随机选择出来的。如果这个投标被接受而且是有利可图的，这个策略的价值就被增强并且在类似情况下再次使用它的概率就会增加，但是如果它不成功，再次使用它的机会就会减少(Roth & Erev，1995)。

参与和伴随建模

基于行动者的模型已经成功地运用于第三世界国家的农村地区，以帮助其

对稀缺自然资源如灌溉用水的管理。这种对基于行动者的模型的神奇运用是因为它们符合"参与式"(或参与性)研究方法。除了应用于研究,多行动者模型还被用于对协商和决策制定的支持以及对培训的支持,例如对塞内加尔的农民(D'Aquino, Le Page, Bousquet & Bath, 2003),法国中央高原护林人和农民(Etienne, 2003; Etienne, LePage & Cohen, 2003)以及南太平洋基里巴斯国(Kiribati)环状珊瑚礁上居民(Dray etc. al, 2006)的培训。

这种也被称为伴随建模的方法(Barreteau, 2003; Barreteau, Bousquet & Attonaty, 2001; Barreteau, Le Page & D'Aquino, 2003)涉及建立一个多行动者系统,这个系统与从在场的人中选出的被调查者有密切联系。作为一个开端,被调查者可能会被问及他们对情况的理解,然后他们会参加一个角色扮演游戏。最终,当获得充足的知识之后,一个计算机模型就被建立起来并与参与者一起用做培训援助或协商支持,回答关于"如果怎样……将会怎样"这种关于可能决定的问题。

例如,为了理解为什么塞内加尔河谷的灌溉计划产生了令人失望的结果,Barrreteau等人(2001)描述了参与式模型的使用。他们开发了一个角色扮演游戏(RPG)和一个被称为SHADOC多行动者的系统,用以代表卷入将水资源分配到灌溉区域可耕种地区的决策制定过程中不同利益相关者之间的互动。在这个例子中,多行动者模型首先被提出来,然后它的主要元素转变为一个RPG(其中表演者等于多行动者模型中的行动者),这一做法的部分原因是为了使基于行动者模型有效,部分原因是为了RPG更容易应用于农村环境中。作者总结这一方法的价值在于"加强了游戏参与者之间的讨论",以及使"在田野中遇到的问题和每个个体分别知道的问题都转变为常识"。

基于行动者的模型的特征

这些被选择出来的例子,在表明基于行动者的模型目前的应用范围的同时,也为说明基于行动者的模型的一些特征提供了例证(Fagiolo, Windrum & Moneta, 2006)。

本体论对应

模型中的计算机行动者与真实世界中的行动者存在一个直接的对应，这使得设计模型和解释它的结果比诸如基于公式的模型更加容易。例如，一个商业组织的模型能包含雇员、客户、供应商和任何其他的重要行动者。在每种情况下，模型可能包含一个行动者代表整个阶级（如雇员），或者可能是每个雇员分别对应一个行动者，这取决于雇员之间差异的重要性。上面描述的电力市场模型中，市场上的每一个主要成员都有一个行动者与之对应。

异质的行动者

一般而言，经济学和组织科学中的理论都作出简化假设，即认为所有行动者在最重要的方面都是相同或相似的。比如，它们处理的是"典型的公司"或在经济上理性的决策者。行动者可能在偏好上有差异，但是遵循不同行动准则的行动者是不常见的，即使当这种情况被允许时，这种具有各自行为方式的行动者的数量可能也很少。因此，有充足的理由认为，除非行动者是同质的，否则分析性的解决方案是非常难或不可能找到的。计算机模型避免了这一局限：每个行动者都能根据他自己的偏好或者依据他自己的行动准则来行事。在供给链模型中能找到这样一个例子，在这个例子中每个商家都能够拥有它自己控制库存的策略。

环境的代表

在基于行动者的模型中，代表行动者直接行动于其中的"环境"是可能的。这可能包括物理的方面（如行动者必须克服的物理障碍和地理障碍），周围位置上其他行动者的影响，以及诸如人群和资源损耗等因素的影响。例如，Gimblett (2002)和他的同事模型化了在加利福尼亚内华达山脉中背包客的运动来检查帮助维持这片荒野保护区的管理政策的效果。被模拟为在这片区域中进行徒步旅行的行动者与模型化区域地形的地理信息系统连接在一起。环境在前面章节中提到的工业区模型中也扮演了重要角色。

行动者互动

基于行动者建模的一个重要好处在于它能够模拟行动者之间的互动。简单来说，这些互动包括从一个行动者到另一个行动者的数据传输，尤其是当另一个行动者就在模拟环境中的附近位置时。如果合适的话，互动也可能会更复杂一些，涉及由某些语言构成的消息的传递，伴之以行动者建构的“表达”和其他行动者对它的解释(但并不一定会获得与述说者所想要表达的相同的意思)。民意动态模型(第1章第2节)是说明基于行动者的模型中互动重要性的一个很好的例子。

受限制的理性

许多模型都隐含地假设它们模型化的个体都是理性的，即他们都是根据一套合理的准则去行动以最大化他们的效用或福利。(另一种替代的方法是将行动者模型化为随机或非理性的行动，即以一种并不能最大化他们福祉的方式去行动。两者在一些模型中都有一席之地。)一些经济学家，特别是那些使用理性选择理论的经济学家，被人们指责假设的个体“过度理性”，即人们为了选择最优的行动过程而经历复杂推理的长链，甚或人们能够遵循无限扩展的逻辑链。Herbert Simon(1957)与其他人都批评这是不现实的，并提出人们应该被模型化为具有有限理性，即他们在认知能力上是受限制的，因此他们最大化其效用的程度也是有限的(Kahneman, 2003)。基于行动者的模型使创建有限理性的行动者变得比较容易。然而实际上，挑战通常并不是限制行动者的理性，而是将他们的智力扩展到一定的水平，在这个水平上行动者能够作出老练复杂的决定，就如同人们普遍会作出的决定一样。[1]股票市场模型和在第1章第2节中介绍的隔离模型都是行动者被设计为只拥有严格限制性理性的例子。

学习

基于行动者的模型能够在个人和群体层次上模拟学习。例如，上面描述的创新网络模型中的公司能够学习如何生产更畅销和更有利可图的产品，并且随

着时间的推移整个部门作为一个整体(即在模型中的所有行动者成为一个集合体)也在学习哪些产品可以形成一个具有兼容性的组合,这样的话一个公司中的产品就能为另一个公司提供购买的组件。学习可以通过以下 3 种或其中任一种方法模型化出来:(1)作为个人的学习,行动者从他们自己的经验中学习;(2)作为进化的学习,行动者群体因为某些行动者的"消亡"和被更好的行动者所取替而进行学习,从而导致群体平均水平的提高;(3)社会学习,一些行动者模仿或被其他行动者教育,从而引起在个体层次上积累的但在整个群体中传播的经验得到分享(Gilbert et al, 2006)。上面总结的创新网络模型是一个含有学习模型的范例:作为个体的创新型公司学习如何制造更好的产品,因为业绩欠佳的公司倒闭并被更好的新兴公司取代,所以这个部门作为一个整体能够进行学习以改善它的表现。

某些用于设计包含学习模型的技术将在下文中进行讨论。

其他相关建模方法

前面一节回顾了一些基于行动者模型已经发挥作用的领域。然而,基于行动者的模型并不适合于每一个建模任务。在开始新的项目之前,值得关注一下其他的替代性方法。这一部分介绍两类应用于社会科学中与基于行动者建模不相上下的建模方式:微观模拟和系统动力学。

微观模拟

微观模拟从描述以个人、家庭或组织为样本的大型数据库开始,然后运用一些准则去更新样本成员,好像时间在前进一样。例如,数据库可能来源于一个具有代表性的全国家庭调查,包含了诸如家庭成员年龄、性别、教育水平、收入、职业和养老金安排等变量的数据。这些数据可能与调查进行时的特定时间段相关。微观模拟允许人们去调查样本将来的状况。例如,人们可能想知道在未来 5 年内样本中有多少人会退休以及这将如何影响收入的分配。如果我们拥有某些关于在一年中个人境况可能发生变化的规律,那么这些规律可以应用于

样本中的每一个人以发现在调查结束后的第1年年末可能发生的改变是什么。然后同样的规律也能重新应用于第2年之后样本出现的状态，诸如此类。当这个老化过程开始进行时，样本整体的合计统计值可以被计算出来（例如收入分布的均值和方差，这可以与调查当时的分布相比较），并能作出关于样本所在的总体中期望变化的推论（Gupta & Kapur，2000；Harding，1996；Mitton，Sutherland & Weeks，2000；Orcutt，Merz & Quinke，1986；Redmond，Sutherland & Wilson，1998）。

微观模拟已经被用来评估在社会安全、个人税收和养老金变化上的分布含义。例如，它有助于估计收入门槛降到国家补助可以支付的水平之下这一变化的影响（Brown & Harding，2002）。试验原型在有数据的情况下也已经被发展出来，它们不仅描述个人也描述公司，并且其中的老化过程不仅受个人特征的影响也受宏观经济变量，如通货膨胀和国内生产总值（GDP）的影响；相关文献参见 http://www.microsimulation.org/。

微观模拟模型的一个优势在于它们不是从一些假设的或随机生成的一组行动者开始，而是从真实的样本开始，像抽样调查描述的那样。因此，与基于行动者模型相比，从微观模拟的结果中读取信息去预测关于真实总体的未来状态是相对容易的。但是也存在两个主要的不足。首先，老化过程需要非常详细的转变矩阵以明确说明目前在某种状态的行动者在下一年将转变到其他某种状态的可能性。例如，需要知道目前在职的人在一年后变成失业状态的可能性。另外，因为这一转变的可能性在男人和女人之间、在有孩子的女性和没有孩子的女性之间、在年轻人和老年人之间等等几乎肯定会不同，所以需要的不是一个单一概率值而是一个条件概率矩阵，这个条件概率矩阵将对应个人状况的每一种组合。获得这样的转移矩阵的可靠估计可能非常困难，需要从非常大量的数据中得到估计。第二，每个行动者都单独地变老并且被看做好像是与世隔绝。微观模拟没有将行动者之间的任何互动考虑进去，而且通常没有空间或地理的概念。举例来讲，有发现认为，如果一个人所居住地区的失业率很高，那么这个人失业后找到工作的机会就会更低，但微观模拟很难将这种发现考虑进去。

系统动态

在用系统动态方法去模型化时，人们创建了一个模型来表达变量间暂时的原因—结果关系，但行动者并不被直接呈现出来。最早和最知名的例子是 Forrester 的世界模型，这个模型被用来预测未来的人口水平、增加的污染和对自然资源的消费率(Forrester, 1971)。系统动态学，正如其名称所体现的，能够模型化交互的变量系统以及能够直接处理因果联系，诸如人口的增加导致资源消耗的增长和反馈回路，就像人口增长依赖粮食供给但粮食供给又依赖人口水平一样(Sterman, 2000)。

通常用图表来呈现系统动力模型比较方便，其中箭头表示变量间的因果联系。图 1.2 给出了一个虽然简单、但非常典型的生态系统模型，其中，羊的繁殖与它们的总数成比例，狼吃羊，但如果羊太少的话，狼就会挨饿。矩形框代表羊和狼的存量，像水龙头一样的符号是存量的流入和流出，菱形方块是控制流动率的变量。羊的总数随着羊的出生而增加。而这发生的比率是由恒定不变的羊出生率所决定的。这个图表明羊以一定的比率死亡，这个比率是活着的羊数

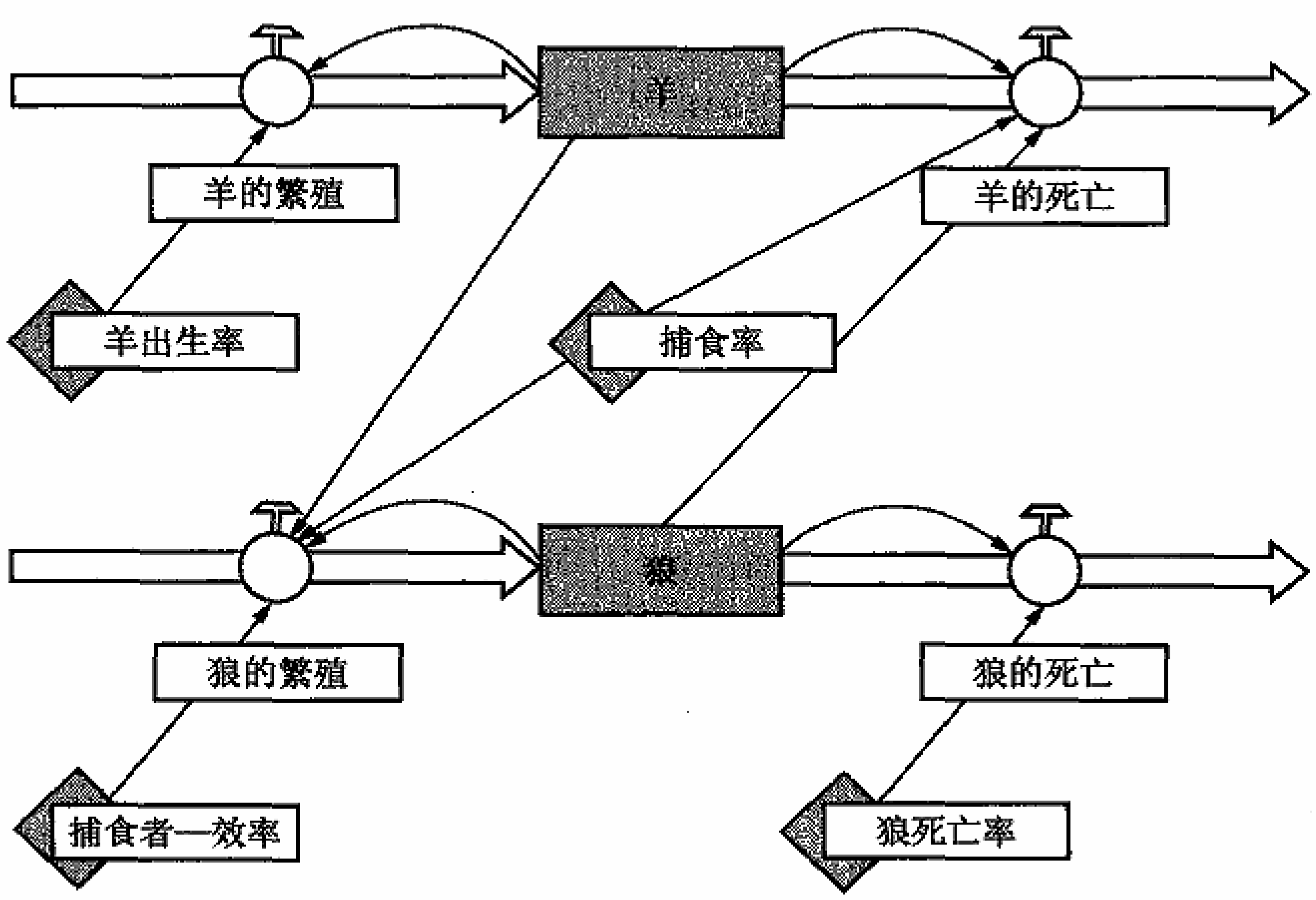


图 1.2　生态系统模型

量(从羊圈符号到标签为"羊—死亡"的流动控制符号的弯曲箭头)、狼捕获羊的概率(从"捕食率"变量中出来的箭头)和狼的数量(从"狼群"中出来的箭头)的函数。虽然这个演示模型关注的是一些假想的狼和羊,但相似的模型也能为有社会学趣味的议题建构出来,诸如违禁药物使用者和缉毒人员的数量,以及公共健康流行病等(Homer & Hirsch, 2006; Jacobsen, Bronson & Operation Research Society of America, 1985)。

系统动力学是基于对联立的微分或差分方程的评估。每个方程在给定其他变量的数值以及目前时间段因果变量的条件下计算一个变量在下一个时间阶段的值。像 Stella(http://www.iseesystems.com)和 NetLogo(http://ccl.northwestern.edu/netlogo/,在第 3 章中将介绍更具体的内容)这样的软件能够帮助绘制图表并且也能够通过运算这些方程来执行模拟。

与基于行动者的模型相比,系统动力学方法处理的是一种集合的而非个体的行动者。例如,在狼和羊模型中,模拟程序将计算出在每一个时间段羊的总数,但每个个体的羊并不会被呈现出来。这使它很难模型化行动者之间的异质性;虽然在原则上,每一个不同种类的行动者都可以有一个独特的种群(例如白羊的种群,黑羊的种群,杂色羊的种群等等),但是在实际情况中如果多于一些不同类型时这个模型就会变得极其麻烦。在一个系统动态学模型中也很难呈现出行动者基于过去经验、记忆或学习的行动。另一方面,因为它们处理的是集合体,因此系统动态学方法适用于有关行为相似的行动者群体的议题。因此,系统动力学对于 Forrester 的全球经济模型而言是一个合适的方法,因为在这个模型中个体的行为是不重要的,焦点在于作为整体的世界状态。

第 2 章 | 行动者、环境和时间尺度

在前一章中我们注意到基于行动者的模型是由一组在环境中行动的行动者构成。在这一章中，我们将开始展示基于行动者的模型被创建和编程的原则。

行动者

行动者通常被习惯性地描述为具有 4 个重要特征（Wooldridge & Jennings，1995）。

（1）独立自主。不存在一个全局性的控制者命令行动者去做什么；它可以在它目前的状态下做任何被程序设定的事情；

（2）社会能力。它能够与其他行动者互动；

（3）反应性。它能够对来自它所处环境的刺激作出适当的反应；

（4）主动性。它有自己会主动追求的一个或多个目标。

然而，这些特征在设计一个行动者时并不是非常有帮助，而且它需要很多想象力去看在大部分模型中拥有全部这些特征的行动者。描述行动者的一个更有效的方法是说它们具有如下的特征：

（1）感知。它们能感知它们的环境，这环境可能包括附近其他行动者的存在。在编程的术语里，这意味着行动者有决定位于它们周围的对象和行动者是什么的方法。

（2）表现。它们有一套能够进行表现的行为。通常，这些行动包括如下的内容：

a. 移动。它们能在一定空间(环境)里移动;

b. 交流。它们能够发送消息给其他行动者并能从其他行动者那里接收消息;

c. 行动。它们能与环境互动,例如,拿起"食物"。

(3) 记忆。它们具有记忆,能够记录它们对于以前状态和行动的感知。

(4) 政策。它们有一套准则,探试法或策略,在给定它们的目前状态和它们的历史的条件下,这些决定了它们现在将要进行的行动。

这些特征能以许多不同的方式被实现。在这一节中,我们描述3个常用的方法:建立一个模拟并且它的行动者直接使用一种编程语言、运用一个生产准则系统、运用神经网络。

特定的编程

几乎所有的基于行动者的模型都是使用一种面向对象的编程语言建立起来的,如 Java, C++,甚或 Visual Basic。这本书不能教授你面向对象的编程(有许多不错的介绍,包括 Eckel, 2005; Niemeyer & Kundsen, 2005),但是由于面向对象的编程构想对于基于行动者的模型而言非常重要,因此这里对它的主要特征做一个简短的介绍。

面向对象的编程将程序发展为对象的集合,每个对象都有它自己能做的一套事情。对象能够在它自己的属性里面存储数据,能够传送消息给其他行动者,并且有决定如何处理数据的方法。对象导向的一般编程方法的优点在于它提供了一个高水平的模块性:例如,有关对象的方法是如何运作的细节能够在不干扰其他程序的情况下被改变。对于基于行动者的模型而言,另一个优点在于行动者和对象概念之间有亲和力:很自然地将每个行动者编程为一个对象。

在面向对象的编程中存在一些基本的概念:

(1) 类别是对象的一个抽象说明,包括对象的属性和它的方法。例如,程序中可能包含了一个被称为"公司"的类别,用以代表在一个经济体模型中的一家公司。这个公司可能具有诸如资本化,员工数量和销售产品种类等属性。它也可能具有一些方法用以描述卖出公司产品给消费者和从供应商处购买公司原

材料的过程,供应商可能是模型中的其他公司。类别可能被专门化以形成更多具体的类别。例如,一般的公司类别能被专门化为制造公司和销售公司。每个专门化类别继承了比它抽象的类别的属性和方法,而且可能增加了新的属性和方法或覆盖了它继承的属性和方法。比如,制造类别可能需要能决定产品数量的方法,产品数量能从原料的单位数量中计算出来;但销售类别可能就不需要这种方法了。

(2) 当程序运行时,类别被具体化以形成对象。例如,生产类别可能被具体化,产生两个对象,一个表示为 XYZ 公司,另一个表示为 ABC 公司。虽然这两个对象有相同的方法和相同的属性,但他们的属性值(如,他们的资本化、规模、产品类型)可能是不同的。一个对象的方法可能向另一个对象传送消息,因此影响到它的状态。例如,XYZ 对象的一种方法可能向 ABC 传送一个消息要求它卖出一些产品,ABC 的一种方法可能以它希望卖出产品的数量和价格作为回应。

正如这个简要的总结应该表明的,从对象导向到基于行动者建模只是一小步。人们为每一类型的行动者建立一个类别,提供保存行动者目前状态(记忆)的属性,并加入合适的方法来观测行动者的环境(感知)和依据某些准则(政策)来执行行动者的行动(表现)。另外,人们需要编出一个调度程序用以在模拟开始时具体化所需要的行动者数目,并给每个行动者一个行动的开始。第 3 章包含了一个建立这一模拟的例子。

生产准则系统

我们在前面一节已经注意到,在大多数模型中行动者都需要具备能力去感知他们的环境状态,从其他行动者处接受消息,依据他们目前的状态积极地选择行动去表现,并向其他行动者发出消息。实现这些的一种方法是赋予行动者以下的内容:

(1) 一组行为准则。这些准则决定行动者将要做什么。依据行动者目前的状态,一个或更多的准则将从这组准则中被选择出来。这些准则通常被称为“条件—行动”准则,因为它们同时包含了条件要素(如果准则要被使用,什么必

须是真实的)和行动要素(为执行准则将要做什么)。比如,一个此类的准则可能是"如果我能在附近找到食物,那么我将向它移动一步"。

(2) 一个工作记忆。这将由储存了行动者目前状态的变量组成。例如,工作记忆可能储存了行动者目前的方位和它目前的能量储备。

(3) 一个准则翻译程序。这是一些程序编码,它们使用工作记忆以便选择哪个准则应该被激活或被"解雇"。它可能需要处理由多于一条准则构成的条件要素被满足的情况,因此需要一些在准则间进行选择的方法。

(4) 一个输入过程。这将从环境中收集消息和感知,并将它们存储在工作记忆中以通过准则处理它们。

(5) 一个输出过程。这将传递消息到环境中,在途中传递到作为预期接收者的行动者那里。

这种安排被称为生产系统,准则为生产准则(Nilsson, 1998; Waterman & Hayes-Roth, 1978)。在生产系统的要素与以上列明的行动者值得拥有的特征之间的对应应该是很清楚的。一个相对简单的生产系统能从诸如 JESS(the Java Expert System Shell, http://www.jessrules.com/)这样的工具包中被建立起来(Friedman-Hill, 2003)。作为替代性选择,也有一些更加精致的系统,这些系统是基于心理学上可能的人类认知模型,比如 Soar(Laird, Newell & Rosenbloom, 1987; Wray & Jones, 2006; Ye & Carley, 1995), CLARION(Sun, 2006)和 ACT-R(Taatgen, Lebiere & Anderson, 2006)。

神经网络

另一种提供行动者必备特征的方法是使用模拟神经网络(ANNs)。它们的设计大致基于生物神经细胞,例如那些在大脑中负责学习或模式配对的神经细胞。ANNs 能够用硬件建立起来,但在一般情况下它们是软件程序或文件库。一个 ANN 由排列在堆积层次上的一些(一般在 10 到 100 之间)单元所组成(Engelbrecht, 2002; Garson, 1998)。每个单元与层级中在它之前和之后的所有单元相连(图 2.1)。每一个连接都从一层中的一个单元向下一层中的一个单元传递信号(在软件执行中,有一个数字代表信号的大小)。每个单元都将来自

前一层的所有信号接收下来，并将每个信号都乘以一个连接权重，将乘积加总，对总和应用一些非线性数学转换（如 logistic 或双曲正切），然后再将信号传送给下一层级的所有单元。第一层的单元从网络之外接受输入，每个单元接受一个，最后一层则发送一个由这一层每个单元发出的信号组成的输出结果。

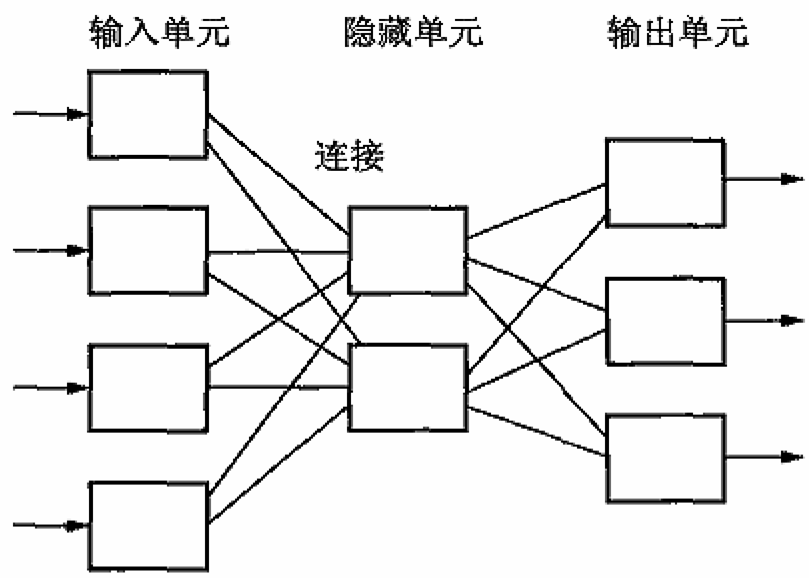


图 2.1 虚拟神经网络单元示意图

在 ANN 最简单的应用中，网络被用来识别各种模式，每当一个特定模式被应用于输入时，就输出一个特定结果。例如，一个 ANN 可能被训练去识别条形码，并输出与条形码线的每个模式相对应的数字。在训练之前，所有的连接权重都被设定为随机数值。将一个条形码作为输入应用到第一层只导致在输出层上一个随机信号模式。一种被称为反向传播算法的方法接着被使用。这种方法将实际输出与期望输出相比较（比如，它将未受训练的网络产生的随机数字与条形码被期望产生的数字串相比较），并调整连接权重以使输出结果更靠近期望输出结果一点儿。然后，另一个输入（如，另一个条形码）被应用，将实际输出与期望输出进行比较，连接权重被再次调整，这个训练过程一直继续到 ANN 在它的认知任务上表现得足够好。这涉及成千上万的训练例子，而一旦网络经历了合适的训练，它就能对它以前从未见过的输入进行解码，或被应用于模型化社会群体理论或金融市场理论（如：Beltratti, Margarita & Terna, 1996; Kitts, Macy & Flache, 1999）。

另一种选择已经在基于行动者的模型中被使用（Acerbi & Parisi, 2006;

Cecconi & Parisi, 1998),那就是创建一大群 ANN(每一个 ANN 对应一个行动者)。这些行动者都被一组随机的连接权重初始化,并被设置一个诸如从一片区域中收集"食物"的任务。行动者对它前面是否有食物的感知被反馈到 ANN 的输入中,输出与行动者的行动诸如移动和摄取食物相接。行动者被赋予一个初始的能量值,在每个时间段上都会有其中的一些能量被消耗掉。如果能量降为 0,行动者就会"死亡",并从模拟中移除出去。然而,行动者可以通过摄取食物来增强能量,食物散布在地理区域当中。因为行动者的 ANN 被初始化时所用的随机连接权重,大多数行动者在寻找和摄取食物时都将失败并会很快死亡,即使有一些会成功。这些更成功的行动者进行再生产,将与他们自己相似的连接权重给予他们的后代,但会有轻微的变异。逐渐地,行动者群体将学会获取食物的行为。虽然一些差异仍将存在,但每个行动者的能力将一代一代地提高。因此,不像反向传播方法那样直接地改变连接权重,这些模型运用了一种达尔文选择的形式,以更具适应性的权重来产生行动者。

与运用生产准则对行动者进行编程相比,神经网络的优点在于它们更具灵活性并且能够发展出研究者没有考虑到的行为。神经网络也被认为比生产准则系统更接近于人类大脑(但必须承认虚拟的神经网络与大脑中的生物神经网络的运作方式非常不同,同时也更加简单)。另一方面,一旦神经网络被训练,就几乎不可能去发现它为什么会运作:人们所全部掌握的就只是一个连接权重的矩阵,这个矩阵不能用单元的活动提供关于其行为的全面解释。但用生产系统,人们可以利用准则去解释行动者的行为。此外,使用通过一个进化过程来训练的神经网络模型是非常占用运算资源的。为了进化一个运行着的行动者,人们需要一个庞大的初始群体和长时间的发展。

环境

前面一节已经描述了构建行动者的 3 种方法。行动者在一个环境中行动,环境提供了行动者之间沟通的渠道,同时也可能包含无反应的对象,比如障碍物或能量源。

通过环境来为行动者之间的全部交流安排路径是方便的，这不仅是因为完成它是一个“自然”的方式，对应于环境在人类事物中的角色，同时也因为它使监督行动者变得更加容易。它也意味着从一个行动者传递到另一个行动者的消息能够被暂时地储存在环境之中(缓存的)，减小了模拟结果依赖于行动者编码被执行的命令偶然性的可能性。当进行缓冲时，来自行动者的消息被储存在一个临时变量中，直到模拟中的所有行动者都有他们轮到的机会。然后，这些存储的信息被传递给接收者。

在许多模型中，环境会包含被动的对象，比如区域的栅栏，行动者可能经过的“路”和连接，为行动者提供能量或食物的资源，等等。这些都可以用与行动者类似的方法来编程，甚至可以更简单，因为它们不需要具有任何能力来对它们周围的环境作出反应。

因为环境对编程而言经常是一目了然的，所以它很容易被忽略；然而在真实世界中环境的影响通常是非常重要的。人类生活的许多复杂性都来源于要处理复杂的环境。我们经常把环境当做一个记忆体(比如，将物体放在特殊位置上以提醒我们需要采取的行动)，当做价值的储存器(比如，货币和其他形式的财富)，以及当做一种使某些行为更加简单的技术性帮助(比如：提供通讯和交通服务的设备)。然而，研究者在设计模拟模型时很少能认识到这种环境复杂性。

随机性

在进行模型设计时，从真实世界的复杂性中进行抽象的一种方法是键入一些随机性。例如，在一个产业网络模型中(第 1 章第 2 节)，有人可能想从一个特定工业部门中发现的公司和公司内部联系中进行抽象以便提出一个关于公司间网络的一般模型。然而，问题是哪些联系应该被建构进模型中？一种答案是随机选择几对公司并在这几对公司之间创建连接。这可以通过许多种方法来完成，主要取决于联系的预期分布。例如，一个公司被连接到另一个公司的机会对于所有公司而言是相同的(随机网络)，这种机会可能与已经被连接到其他公司的公司数量有关(偏好连接)(Barabási，2003；Barabási & Albert，1999)，

或者这种机会可能已经被安排好，所以这些连接形成一个由紧密联系的公司组成的当地群落，在群落之间有一些远距离的联系（“小世界”网络）（Watts, 1999; Watts & Strogatz, 1998）。这些替代性方案为网络提供了不同的结构特征，这些结构特征可能会导致相当不同的行为（Pujol, Flache, Delgado & Sangüesa, 2005）。

随机性可能也被用于模型化沟通错误和“噪音”的影响。例如，Axelrod（1997b; Axelrod & Dawkins, 1990）提出了一个具有影响力的关于文化传播（Axelrod, 1997c）的模型以解释如果进行互动的人们的信仰和态度倾向于变得越来越像，为什么差异最终仍然不会全部消失。他的模型展现了具有同质性文化特征的稳定区域的出现，这些同质性文化特征有方言、民族主义信仰、宗教风俗等。在他的模型中，行动者拥有“标签”，一组包含 5 个数字，用以描述他们在 5 个文化特征或维度上的文化特点。[2] 两个行动者进行互动的机会取决于他们标签的相似性。如果他们进行互动，5 个特征中的 1 个特征就会被随机选出，一个行动者接受与其互动的另一个行动者在这个特征上的值。经过几千次互动之后，独特的区域出现了，在此区域中所有的行动者共享相同的特点，并且与其他区域中的行动者没有共同的特点。Axelrod 评论说他的模型显示了局部的收敛如何能够引发全球的极化。然而，稍后的研究（Klemm, Eguíluz, Toral & San Miguel, 2003）显示了这种行为严重依赖于行动者精确地复制其他行动者文化特点的值，并且在行动者的特点中也不存在“文化漂移”。相反，如果行动者接受了其他行动者特点的略微变形的复制，或者行动者的信念偶尔发生了随机变化，那么一种单一文化就可能出现，而不是出现许多独特的区域，每个区域有不同的文化。这就是一个关于随机变化或“噪音”如何能够使一个模型的结果产生根本性差异的例子。

时间

在大多数情况下，模拟过程好像是由一个时钟精心安排进行的，一下又一下。每一下，所有行动者都被给予一个转变。因此，时间被模型化在离散的时

间阶段中。每一个时间段都持续相同的模拟期。模拟从零时间点开始，持续足够长的时间，直到所有的行动者都“死亡”或停止行动为止。

当使用离散的时间段来设计基于行动者模型时需要考虑 3 个关于时间的问题：

(1) 共时性。在前一节中我们已经提到需要注意消息从一个行动者发送到另一个行动者的时间。比如，如果行动者 A 发送一个消息给行动者 B，B 回复，然后 C 再发送一个消息给 B，这个结果可能与 A 发送消息给 B，接着 C 发送一个消息给 B，然后 B 再发送一个消息给 A(例如以“内部者交易”模型为背景来思考这些顺序)的结果截然不同。这是关于行动者调用次序这种更普遍性问题的例子(Huberman & Glance，1993)。如果使用一台普通的电脑，因为一次只能做一件事情，因此，实际上行动者不能同时参与行动。处理这一情况的 3 种可能方法如下：

a. 以相继顺序调用每个行动者(行动者 1，行动者 2，行动者 3，行动者 4，行动者 1，行动者 2，行动者 3，行动者 4，…)(有序不同步执行)。这通常不是一个好的解决方案，因为模型的表现可能极大地受到所使用次序的影响。

b. 在每个时间段上都以一个不同的随机顺序调用每个行动者(随机不同步执行)。这种方法的优点在于可以通过多次运行模拟，每次以一个不同的次序，来考察次序的影响。

c. 以任何方便的顺序调用每个行动者，但用行动者所在的环境来缓冲所有的互动，因此在所有结果输出之前对行动者的所有输入都已完成(模拟的同步执行)。如果这种方法能够实现的话，它就是最佳选择，虽然其安排比较复杂，并且有时候模型的要求也会妨碍它的使用。

(2) 事件驱动的模拟。这 3 种安排行动者调用顺序的方法假设所有行动者在每个时间段里都需要有一次行动的机会，尽管实际上一些行动者在他们的时间档期内可能什么也没有做。另一种不同的方法是：使用一种事件驱动的设计，在这个设计中只有那些需要行动的行动者才被调用。事件驱动模拟的构想在于不认为有一个恒定不变持续期的时间段，模拟是从一个事件跳跃到另一个事件。模拟的“时钟”一直在被上发条直到下一个事件发生的时间到了。例如，

假设我们正在设计一个关于组织决策制定的模拟并且有一个模型，其中的行动者是决策制定委员会。模型的焦点为每个委员会作出的决定以及这些决定如何从一个委员会传递到另一个委员会。在委员会会议之间发生了什么是不用考虑的，而且会议本身被认为是瞬时发生的。在这个模型里，具有一个规则的时间步骤是无效的，因为在大多数的时间段里没有有趣的事情发生。相反，应该设计一个程序便于从一个委员会会议时间“跳跃”到下一个会议时间。

（3）校准时间。在规则的和事件驱动的模拟模式中，经常有将模拟时间与真实时间匹配的问题。例如，如果有一个消费者行为模型，在这个模型中某人想要研究消费者对一个新产品介绍的反应。比较有趣的事情在于大部分消费者接受这个新产品所要花费的时间。这个模型可能会显示这需要花费多少时间段，但是人们如何能将这转换成真实时间中的星期和月份？一个解决办法就是在现实中观察这个过程，并将新产品接受趋势的形状与模拟的结果相匹配，但这只是给出了大致的解决办法。另外，在这个和其他例子中，关于将什么当做模拟和真实世界中的起始时间或零时间点仍然是一个难题。虽然模拟是在一个界定清晰的时刻开始的，但是在真实世界中，很少有什么活动（例如，一个新产品上市）是在一个清晰的时刻准时开始的。这些问题都需要注意；不存在一个永远适用的一般性解决方案。

第 3 章 | 在社会科学研究中使用基于行动者的模型

在过去的十几年里，基于行动者的模型已经发展出或多或少标准化的研究程序，包括一系列的步骤。像大多数社会科学研究方法一样，这是实际执行程序的理想状态，事实上，许多步骤都是并行发生的，并且整个程序在构想被提炼和发展时被重复地执行。尽管如此，将这些步骤弄清楚以作为进行基于行动者建模研究的一个指南是有帮助的。

在研究的早期阶段，将宽泛的研究议题缩小范围到具体的研究问题是有必要的。研究问题是一个在研究工作里面有一个能回答它的实际机会的问题。如果一个研究问题太模糊或太宽泛，是没有多大作用的，而且研究也会令人失望，因为它不能提供期望的答案。如果错误出在明确性方面通常会更好一些：是太集中而不是太模糊。有时候将界定研究问题想象成剥洋葱皮也会有些帮助，从调查研究的宽泛领域开始，通过特定议题，到一个研究问题，对于这个问题仅仅只需对你所发现的做一个简要陈述就能回答。

正如我们在上面已经注意到的那样，基于行动者的模型习惯去研究的，是那些在社会或宏观层次上的规律性已经被观察到，但关键是这些规律性如何被解释的问题。经济学家经常称这些规律性为程式化事实(Kaldor，1961)。例如，在第 1 章中描述的 Schelling 模型从观察到邻居的种族隔离开始并寻求通过模型化单个家庭户的决策来解释它。同样，在第 1 章中描述的电力市场模型也力图通过供应商的动机来解释(和预测)电力供应模式和市场价格。

在明确了研究问题并确认了有待解释的宏观层次的规律性之后，下一步就是要明确模型中将要涉及的行动者。它们可能全部都是一种类型，也有可能是

不同的类型。例如,虽然在第1章中回顾的民意动力学模型只涉及一种类型的行动者,但那些意见发生改变的行动者也被模拟出来了,第1章中提及的某些工业区域模型涉及一些独特类型的公司。对每一类型的行动者而言,需要将行动者的行为放在不同的环境之中,这些环境通常是作为一组条件—行动准则(参见第2章第1节)。在两个列表形式中来完成这一步是有帮助的:一个列表展现了环境(包括其他行动者)影响行动者的所有不同的方式,一个列表展现了行动者能够影响环境(再次包含了其他行动者)的所有方式。然后,人们可以写下当行动者必须对环境变化作出回应时的条件,以及当行动者需要对环境作出行动时的条件。然后这些列表可以被提炼用以创建行动者准则来显示行动者如何对环境刺激作出作用和反作用。

在这个阶段,人们将会对模型中需要的行动者类型和它们的行为有一个很好的构想。我们有必要去考虑环境应该采取的形式是什么(例如,是否需要是空间的,在其中行动者有一个确定的位置,或者行动者是否应该被连接进一个网络),以及考虑为了表明模型正在生产所期望的宏观层次的规律性,模型需要展现的特征有什么。一旦所有这些都被考虑过,就可以开始设计和发展形成模拟的程序代码。

在模型已经被建构起来以后,人们就开始了检查它是否正确的漫长过程。这种检查非正式地被称为除错,正式地被称为检验。检验的任务是确保模型满足它意图实现的具体规划。它与有效性验证相当不同,有效性验证是检查模型对于被模拟现象而言是一个好的模型。一个模拟能够满足检验的标准,是因为它可以按照被期望的方式运行,但是如果模型的具体规划是对社会世界中原型的一个拙劣描述,那它就不是一个有效的模型。

在成功的检验之后,就可以着手进行有效性验证。有效性验证的首要标准是检验模型是否展现了研究试图要去解释的宏观层次的规律性。如果它做到了,就可以证明编程进入行动者的互动和行为解释了规律性为什么会出现。然而,必须提防替代性的解释,因为可能存在其他同样的或看似更合理的行动者行为,它也会导致相同的宏观层次的规律性。因此,需要进行灵敏度分析,来检验当模型参数被改变时结果是否也会变化。同样重要的是,要考虑不同的行

动者行为是否能导致相同的结果，特别是一个更简单的模型是否能得到相同的结论(如果得到相同的结果，简单的模型通常应该比复杂的模型更受青睐，使用的原则是简单的解释比复杂的解释好，如果两者在解释上做得同样好的话)。

因此，在探讨了模型的宏观行为之后，接下来就应该将模型的结果与来自社会世界的经验数据做比较。正如我们将会在下文中看到的，模型结果和数据间的这种对比进行起来并不容易，而且经常并不会得到期望中的清晰的答案。大多数的模型都是随机的，即都涉及随机过程，所以人们可能并不清楚模型输出和观察数据之间的差异是因为随机因素所致还是因为一个不好的模型所致。在收集有效和可靠的数据时也经常存在较大的困难，特别当需要与模型输出结果相比较的数据要通过观察很长时期才能收集到时更是如此。

最后，人们可以得出一些结论，期望可以回答开启此过程的研究问题。另外，如果人们对模型有信心的话，可以试验这个模型，从而确认先前不受怀疑的规律性。

发展基于行动者的模型的例子

在这一节中，我们引用一个“集合体”(Gilbert，2006)模拟作为例子来描述发展一个简单的基于行动者的模型的过程。在第 4 章中，我们将看到这个模型是怎样被编程的。

许多相互关联的社会现象是很难被模型化的，甚至很难被描述出来，因为它们的边界是流动的，涉及的人也总是在不断地变化中，而且也不存在被所有参与者都共享的单一特征。一些例子如下：

(1) 青年亚文化圈，诸如“庞克族”(Widdicombe & Wooffitt，1990)或“哥特派”(Hodkinson，2002)；

(2) 科学研究领域或专业(Gilbert，1997)；

(3) 艺术运动，诸如前拉斐尔派或旋涡画派(Mulkay & Turner，1971)；

(4) 社区，诸如伦敦的诺丁山或纽约的布朗克斯(O’Sullivan & Macgill，2005)；

(5) 武装革命或恐怖活动的成员(Goolsby, 2006);

(6) 工业部门,诸如生物技术(Ahrweiler, Pyka & Gilbert, 2004)。

虽然人们可以很容易地说出自己很熟悉的例子,而且尽管这些例子非常普遍并且很容易被辨认出来,但是将某人关于它们的直觉放置于坚实的基础之上却是件很困难的事情。作为开端,还不存在一个普遍被接受的词去命名这个现象。亚文化、领域、社区、专业和运动等术语都被用于特定的类型,但是这些术语中没有一个能恰当地描述它们的全部。一个密切相关的概念是"构型"(Elias, 1939/1969),虽然严格说来这个概念应该只能应用于个人,而不能应用于组织或其他类型的行动者。在这一节中,由于缺少更专业的术语,我们使用词语"集合体"作为通用的术语。需要注意的是,组成集合体的单位可以是人(就像上面大多数例子里的那样),也可以是组织(如生物技术公司)。

获得对集合体的更好理解的第二个障碍在于,根据定义,在它们周围没有明确的边界。这意味着不可能去计算它们的成员,因此也不可能对它们随时间的发展以及它们的发生率等运用更常用的定量分析。

第三,集合体从它们成员的行动中产生的方式并不容易理解。此处发展模型的目的是说明一些关于个人行为(微观基础)看似合理的假设是如何能够产生在宏观层次上可观察到的集合体。

宏观层次规律性

所有集合体在或多或少的程度上都拥有以下这些内容:

(1) 虽然集合体的例子通常在总体层次上比较容易命名和描述,但对其精确的界定被证明是相当棘手的,而且存在商议和争论的可能(举例来讲,有许多不同的地区也能够被描述为诺丁山,范围涉及当地正式的行政区域到与之同名的电影拍摄地点)。

(2) 没有一个已经被大家接受的共识性界定能够被用来从那些"外部"者中区分出"内部"者来(或从非成员中区分出成员来)。例如,虽然从某人的穿着方式,人们可能认为他或她是一个"庞克",通过指出这个人的信仰、行为或熟人,其他人(包括这个人本人)可能会对这个归类提出质疑。这些信仰、行为或熟人

都可能与作出的关于成员身份的决定有选择性关联。特别需要指出的是，不存在一个可观察到的特征是所有“内部”者都具有但同时又是所有“外部”者都不具有的。例如，集合体并不是正式的组织，在正式组织中拥有书面或口头合同的雇员与组织成员相区别；也不是政治党派，在其中最少需要一份正式的支持声明并以此来界定成员；也不是社会阶级，在其中外部确认的客观标准被用于对人们进行分类(典型的是某人的职业)。

(3) 尽管如此，许多成员将共享共同的特点(例如，在一个研究领域的科学家可能有相似的教育背景，进行类似的前期研究，并且相互之间认识，因此即使不涉及技术、理论或研究对象，所有以上这些无一例外也都会在研究领域中涉及)。

(4) 集合体中的成员会被赋予拥有一些相关的知识(例如，专业领域的科学，或在青年文化中被认为是“酷”的事物，或诺丁山当地的地理环境)。然而，没有成员会拥有全部的知识；知识是社会性的。

(5) 被认为与集合体相关的特征的改变。例如，研究者不会无限期地连续研究完全相同的问题；一旦他们解决了一些问题，他们便向新的问题转移，不过仍然会在相同的研究领域。大多数政治运动都会随着时间的推移而改变它们的宣言，以反映他们当前的想法和他们看到的发生在其周围的社会问题，尽管他们仍旧是相同的运动，拥有许多相同的追随者。青年文化总是在不停地变换被认为是流行的事物。

(6) 一些参与其中的人被广泛地认为(比如被其他人认为)与其他人相比他们是更中心的，更具影响力的，有更高地位的或是领导者。例如，一些科学家被认为比其他科学家更加杰出，一些亚文化圈中的成员比其他的成员更“酷”，等等。

微观层次行为

在前一节中提到的集合体的一个共同特征是行动者——即组成集合体的人或组织——都有一些特殊的知识或信仰，尽管这些知识在集合体成员间是社会性分布的。例如，对于科学家而言，是关于他们研究领域的知识；对于青年亚

文化而言，是关于目前流行是什么的知识。因此，并不是每一个成员都有相同的知识，对知识的掌握通常是集合体的一个主要特征(Bourdieu, 1986)。在模型中，我们假设所有个体、成员或非成员，都有相同的知识，但是这一知识是什么在行动者之间是不同的，并且随着时间的推移而变化。我们用这一知识来为行动者定位：行动者某个时刻在一个抽象知识空间中的位置是他/她在那个时刻拥有知识的函数。

第二个假设是一些行动者具有比其他行动者更高的地位，而且所有行动者都被激励试图通过模仿高地位的行动者(通过复制他们的知识)来获得地位。例如，在一个受时尚驱动的集合体中，所有行动者都想要尽可能的时尚，这意味着要接受被他们认为具有最高地位的那些人的服装款式、音乐品位或任何东西(Simmel, 1907)。然而，地位也是稀缺的函数：一个行动者并不能持续拥有高的地位，可能有许多其他的行动者具有非常相似的知识。例如：一个时尚偶像必须始终引领大众；一个科学家只有当他/她的研究与众不同时才能被频繁地引用；一个革命者只有在他/她与“小卒”相比较脱颖而出时才能赢得同事的尊敬。

第三，我们假设最高地位的行动者想要维持地位，但如果他们开始被受他们吸引的追随者排挤，他们就很难维持自己的地位。在这种情况下，我们假设高地位的行动者会被激励去进行创新，也就是在知识空间中去寻找还没有人群的临近位置。

因此，对行动者而言存在两个相互抵消的趋势——一方面，他们想要更接近于时尚；另一方面，他们想要具有排他性并且能够通过改变代表地位水平的位置来实现这种排他性。我们将会看到，处理这一张力在宏观层次上产生了典型的集合体模式。

设计一个模型

相关的模型

有一些一般化的模型来处理相似的问题：

(1) Boid 模型(Reynolds, 1987)的行动者试图与所有其他行动者都维持一个合适的距离，因此表现出随着协调运动而移动。行动者有 3 个转向行为：分

离，以避开附近的行动者；调整，朝与附近行动者平均值的相同方向移动；结合，移向附近行动者的平均位置。这个效应就是行动者像羊群或鱼群一样移动。这些模型展示了使行动者在“张力”中开展行动的影响：例如，分离行动与结合行动之间存在张力。然而，在这些模型中没有寻求地位或创新的观念。

(2) 创新模型(Ahrweiler & Gilbert，1998)的行动者能够根据他们当前的知识来进行学习和行动。行动者也交换知识和创造新的知识。然而，在这些模型中没有关于集合体的具体构想。参与创新的一组行动者是被事先设定好的。

(3) 少数人游戏(Slanina，2000)是来自于众多文献中的一个例子。这个模型也被称为 El Farol 酒吧模型，模型中有行动者想要去酒吧，但只有当一小部分其他行动者也选择去那儿时这些行动者才会去。行动者根据他们自己以前在酒吧中遇到的行动者数量这一经验来作出决定。每个行动者都有一些策略可以用来与他/她最近去酒吧经历的记忆相结合，以对是否要在当前时间段去酒吧作出决定。这些策略根据它们成功与否被评分(当行动者到达酒吧时，是否过于拥挤)，不成功的策略会被放弃。随着时间的推移，能够实现一个动态均衡，酒吧里行动者的数量与行动者用于判断这里有太多行动者的极限值相匹配。在这里，这个模型有某些问题解决导向的特征，但是没有集合体的表征。

模型

集合体模型由一个行动者能够在其上移动的平面组成。这个平面是一个圆环面，其中每一点代表一个特定的知识体或信仰集。因此，行动者并不是在一个物理空间的代表物里运动，而是在“知识空间”里运动。虽然在二维空间中(更准确地说是在一个圆环的平面上)呈现知识空间可能过于简化，但这是为了更容易可视化。

行动者在知识空间中的运动代表了它在知识上的变化。因此，如果一个行动者模仿另一个行动者，就会看到它在知识空间中朝那个行动者移动。反之，如果这个行动者创造和发现了其他行动者不具有的知识，就会看到他离开其他行动者，进入空间中先前空白的区域。

行动者刚开始时在平面上是随机分布的。他们对于自己或其他行动者以前的位置没有记忆。每个行动者都做如下的行动：

(1) 计算与他直接相邻的其他行动者有多少；

(2) 如果行动者的数量超过了极限值，他将转向与邻近的其他行动者移动的大致方向相反的方向，然后移动一个随机的距离；

(3) 如果行动者的数量等于或低于极限值，他环顾四周的位置以找到一个相对满员的区域，然后从他现在所在的位置沿着那个区域中心的方向移动一个随机的距离。

每一个行动者都不同步地行动，无限地重复这一系列的行动。这个算法需要 4 个参数(参见图 3.1)：

(1) 环绕一个行动者的环形区域的半径，在此半径中行动者的数量被计算出来以决定这个行动者是否“拥挤”或“孤单”(局部一半径)；

(2) 行动者数量的极限值，低于它时行动者是“孤单的”，高于它时行动者是“拥挤的”(极限值)；

(3) 环绕一个行动者的环形区域的半径，在此半径中如果行动者是孤单的，计算行动者的数量以找出最大值在哪里；如果行动者是拥挤的，找出行动者运动的一般方向以决定他将要移动的方向(可见一半径)；

(4) 行动者移动的距离。距离是从这个参数的一个统一分布中随机选择出来作为最大值(速度)。

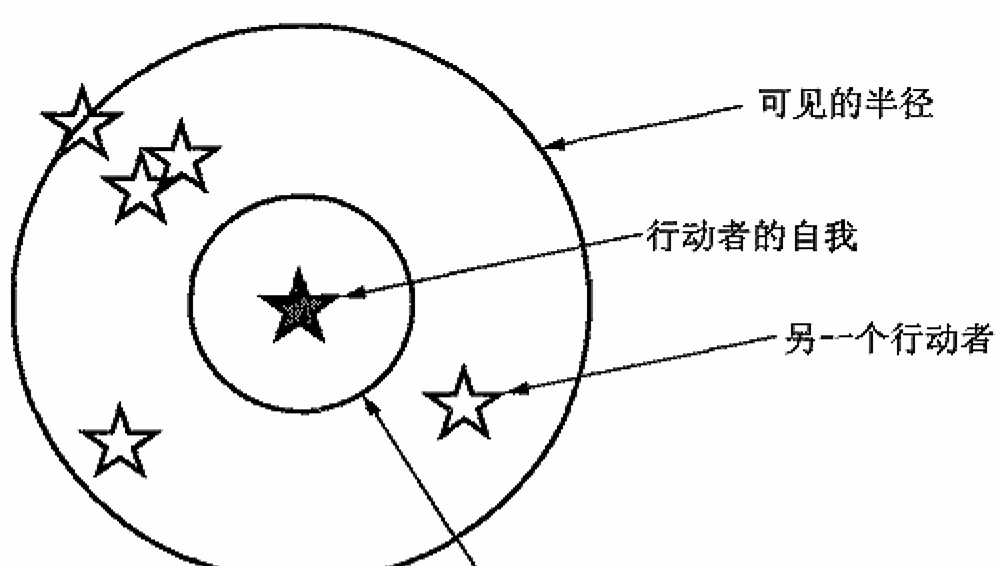


图 3.1 集合体模型的参数

检验

对集合体模型的检验和有效性验证我们会在第 4 章中再做讨论，安排在我们描述完如何对它进行编程之后。首先，我们需要考虑在检验基于行动者的模型时产生的一些普遍性问题。

检验：处理漏洞

你应该假设不管你多么仔细地设计和建构你的模拟，它仍将包含漏洞（做了与你想要的和期望的不同事情的编码）。当你第一次运行一个新的模拟模型时，它很可能有许多漏洞，其中的大部分漏洞都很容易观察到，因为模拟会给出异常的结果或导致程序崩溃。更令人担心的很有可能是，即使当你已经在编码上进行了处理去除了明显的漏洞，但仍有一些漏洞会隐藏着，使你在未察觉时一直被困扰。经验法则把剩下的漏洞数量看做是遵循了一个负指数函数——漏洞数量首先快速下降，然后趋于平稳，但从来不会触及 0。有时甚至已经发表的模拟也会受到漏洞和错误解释的困扰（Edmonds & Hales，2003；Galan & Izquierdo，2005；Rouchier，2003）。

有一些技术可以减少含有漏洞的机会，而且能够使发现它们变得更加容易（Ramanath & Gilbert，2004）。下面列举一些：

（1）精细地编码。当你撰写模拟程序时，仔细稳重地去做，不要匆忙地使编码运行并寻求捷径，这样任何节省下来的时间会在去除漏洞时花费的额外时间中失去。使用对象导向的语言和在你的模型情境中具有意义的变量名称将会起到帮助作用。

（2）包含大量的输出和诊断。如果编码不能给出输出结果以显示当程序运行时发生了什么的话，是很难在这样的编码段落中发现漏洞的。不要满足于只显示模拟的结果，即使在去除漏洞期间，你也需要显示中介值。同时也需要注意决定显示什么以便于你得到好的诊断，但也不能有过多的结果以至于你不能找到漏洞的症状，因为其他输出结果会将它们埋没。

(3) 一步一步地观察模拟。一次运行一条或一个功能的编码，观察变量值、参数值和属性值是如何变化的，并检查它们是否按照期望的方式在发生变化。虽然这种工作是缓慢和乏味的，但是它有助于确保编码按期望运行，至少对于能观察到的运行而言是有帮助的。通常，编程环境提供了使编码更容易被管理的特征。

(4) 增加断言。如果你知道变量必须取某些值而不是取其他值，当模拟运行时检查有效值，并在数值超出范围时显示警告(这样的检查被认为是断言)。例如，如果两个行动者不能占据相同的空间位置，那么在每一个时间段中都对此进行检查并在这一要求被违反时进行报告。

(5) 添加去除漏洞开关。你可能担心如果所有编码都需要辨别漏洞的话，诸如断言和诊断，将使模拟运行变得很慢而令人无法接受。可以在你的程序中包含一个能够设定到去除漏洞水平的全局变量：从 0 到最大值。在每次去除漏洞陈述前都对这个变量进行一次检验，以查看陈述是否应该在目前去除漏洞水平上被运行(在某些语言中达到与条件编译相同的效果是可能的)。

(6) 增加评论并及时更新。所有编程语言都可以插入评论——这是让编程员看的文本，并不会作为程序编码被执行。运用这个特性对每一个功能、程序、方法和对象添加评论。评论应该描述接下来的编码块是什么，以及这个编码块如何运行，但是这种描述只是在概念层次上而不是在执行层次上(即不是重复程序编码，而是说明编码想要实现什么目标)。一条经验法则是评论的行数大约是编码行数的三分之一。评语很容易过时，因为它描述程序以前的情况而不是现在的情况。应该预留一些时间来定期更新评论。在写评论时，假设读者是像你一样能够编程的人员，但对你的模型知之甚少或完全不知(在离开编程几个月之后，这可能是对你本人的一个很好描述，所以不要以没有其他人会看你的编码为借口逃避写评论)。

(7) 使用单元测试。单元测试是一个越来越流行的用以减少漏洞的软件工程技术(Link & Fröhlich, 2003)。它包括写一些测试编码并在写这些编码的同时去运行程序。这个构想就是在较小的相对自我包含的编码块或单元中发展程序。这样一个测试“工具”被创造出来为单元提供一系列的输入并将输出结

果与一列期望结果做对照。然后这个测试工具将自动地运行每一个输入，检查实际产生的期望输出结果。一旦单元通过了测试工具的所有测试，你就可以继续写下一个单元。有时候可能要对第一个单元做一些修改，然后这些修改必须被再次放入测试过程以确保修改没有引入任何新的漏洞。当许多单元都已经被写好时，测试工具会被用来在全部单元上自动执行所有测试，因此可以基本确保漏洞没有作为编码发展的结果在不经意间掺杂进来。当程序发展时，应该写出一些另外的测试以验证多个单元的集合体以及它们之间的互动。

(8) 用已知情境的参数值测试。如果确切地知道了任何情境的参数值和输出结果，就对模型是否复制了期望行为进行测试。这是大多数人首先会用在模型上的测试，但是这个测试本身是一个相当弱的测试，不能有充分的信心保证模拟没有漏洞。

(9) 使用拐角测试。用可能的参数极端值来测试模型并确保输出结果是合理的(这个名字来源于这种参数值标记了一个含有所有可能参数值的参数空间的角落)。例如，可以测试一下当你的模拟在没有任何行动者时运行会发生什么，以及当模型在拥有所允许的最大数量行动者时运行会发生什么。

有效性验证

一旦人们提出了一个基于行动者的模型，很明显就需要检查它的有效性，即实际中它是否是它意图要代表事物的一个好的模型。然而，有效性验证的理论和实践都比人们刚开始的预期要更复杂和更具争议性。这些议题与建模者瞄准的各种目标有关，这些目标暗示了有效性验证的不同标准，以及要求具有足够数量的合适的社会科学数据以允许进行系统有效性验证的艰难(Troitzsch)。我们在讨论有效性验证的一些技术之前需要先思考一些概念性问题。

基于行动者的模型主要用于将理论形式化(例如，Schelling 的居住隔离模型，参见第 1 章)，在这种情况下模型很可能被定位在一个非常抽象的层次上；或者它们旨在描述一类广泛的社会现象，比如工业区的发展或消费者行为；或者它们被希望为一个特殊的社会情境提供一个非常具体的模型，比如在第 1 章中

提到的一些电力市场模型，其中市场的明确特点，包括发电厂的位置和消费者的需求模式，都是相关的。这些类型的基于行动者的模型需要不同的方法去进行有效性验证(Boero & Squazzoni, 2005)。

抽象模型

抽象模型的目标是展示许多社会生活领域背后的一些基本社会过程。一个很好的例子是 Epstein 和 Axtell 开拓性的著作《成长的虚拟社会》(*Growing Artificial Societies*)(Epstein & Axtell, 1996)，它展现了关于一个虚拟社会中经济的一系列连续逐渐复杂的模型。另一个例子是本章先前介绍的集合体模型。这些模型没有想要模型化任何特殊的经验案例，而且对于某些模型而言，可能很难发现其与观察数据有任何紧密的联系。例如，Schelling 的模型一般建立在一个规则的圆环形网格上，行动者被二分为两种类型(如红和绿)。模型的这些特征明显没有想要成为任何真实城市或真实家庭的经验描述。那么这样的模型如何才可能是有效的?

答案是将这些模型看做是理论发展过程的一部分，并将通常应用于评价理论的标准应用于它们。也就是说，抽象模型需要在宏观层次上产生被期望的和可解释的模式；需要基于在微观层次上貌似合理的行动者行为准则；并且最重要的是能够产生进一步更具体的或“中层”的理论(Merton, 1968)。正是这些中层理论和基于它们之上的模型能够根据经验数据进行有效性验证。如果一个抽象模型是使用演绎策略创建的，就会存在一些关于行动者行为和期望的宏观层次模式的假设。因此第一个有效性验证测试就是评估模型是否的确产生了期望的宏观层次模式。更加彻底的测试将会看到当模型的参数被系统地改变时将会发生什么(参见第 3 章第 4 节，关于敏感性分析)。人们可能期望宏观层次模式在参数变化时维持不变，或者期望如果它们变化，这些变化也可以被解释。例如，在 Schelling 的模型中，可以改变行动者的容忍度。当容忍度取较小值时，家庭很少能找到他们感到幸福的地点，并且模拟要花费很长时间才能达到稳定状态，如果它曾经达到过的话。当容忍度取足够大的值时，不管家庭的邻居颜色如何，他们都会感到满意，并且最初的居住随机分布模式也很难改变。

一旦这些基本的测试都已经通过了，就能够评估模型是否能够被用来向理论报告具体的社会现象，然后就能够测试那些理论了。例如，Schelling 的模型，虽然是针对种族隔离问题提出来的，但却比这一问题更具普遍性和抽象性。它能被应用于行动者的任何角色。但是当被用于解释种族隔离时，人们就需要开始发展这一理论以将其他因素包含进来，这些因素对于决定城市中的处所位置无疑是重要的，包括支付能力和房屋储备的可得性，存在两个以上的种族群体，存在不属于任何种族或属于不止一个种族的人群，以及种族态度的功能形式。(例如，Bruch 和 Mare 2006 年指出，Schelling 模型中的隔离效应取决于行动者对幸福或者不幸福有一种明确的态度，如果行动者在从幸福到不幸福的范围内有一个平缓连续的态度，就不会导致群体聚集的出现。)

中层模型

我们在第 1 章中提到的那些模拟消费者行为，产业区域，以及创新网络的模型等，都被认为是"中层"模拟：它们都意在描述一个特定社会现象的特点，只是以一种具有足够一般性的方式使它们的结论能够被广泛地应用于大部分的工业区域中而非仅此一个。

这种模型的普遍性特征意味着通常不太可能非常精确地将它们的行为与任何特定的可观察到的例子做比较。相反，人们期望满足于质上的相似性。这意味着模型的动力学应该与观察到的动力学相似，并且模拟的结果应该能显示与真实世界观察到的相同或相似的"统计信号"；也就是说，两者结果的分布应该在形状上相似(Moss，2002)。

例如，人们在创新网络中找到的公司与在相同产业部门中的其他公司有合作性联系。如果计算每个公司合作伙伴的数目，并且绘制出合作伙伴数目的对数比上与这些合作伙伴有联系的公司数量的对数的图，这个图大致就是一条有固定斜率的直线(如，Powell，White，Koput & Owen-Smith，2005，图 3)。在取对数的变量之间的线性关系就是幂次法则的统计信号，而且它是许多社会网络的特点，包括从公共电力网络到万维网(Barabási，2003)。我们可能期望创新网络的模拟也能够展现一个具有类似斜率的关于公司内部联系的幂次法则

分布。

中层模型的一个例子是 Malerba，Nelson，Orsenigo 和 Winter(2001)关于电脑产业的研究。他们描述自己的模型是“历史—友好的”，其意思是尽管此模型不能复制电脑产业的完整历史，但是它能够

> 以一种格式化和简化的方式把握关于电脑产业进化决定因素的具有鉴别力理论的焦点。它能够用与基本理论假设相一致的参数设定来复制出电脑产业历史上的主要事件。

他们也注意到“在参数标准设定上的变化实际上会导致不同的结果，‘另类历史’与可观察到的程式化事实的基本起因相一致”。

传真模型

传真模型旨在为一些特定的原型尽可能精确地提供复制品，通常是想要运用它对原型的未来状态作出预测，或者预测如果某些政策或规则被改变时将会发生什么情况。例如，一个商业公司可能对找出减少发出进货订单之间的时间间隔对他们库存水平的影响感兴趣。这就要求模型能够准确地表示他们所有的供应商，每一个供应商提供的货物，以及这些货物的单位数量等，以便能够作出合理的预测。另一个非常不同的例子是 Dean 等人(1999)对美国西南部阿那萨吉印第安人的研究。大约在公元前 1800 年，这些印第安人在长屋河谷开始种植玉米，但到 3000 年后就放弃了这块区域。Dean 等人提出模型的目的就是回溯河谷中的定居模式并将这一模式与考古学纪录一户一户地进行匹配。

如果能够获得这样精确的匹配，它们就会非常有用，不仅能对作为模型基础的理论有一个强有力的证明，而且也有助于进行看似合理的预测。然而，有理由相信那些对特定现象的观察进行精确匹配的模拟可能是极少的并且可能只限定于相当特殊的情况。大多数的社会模拟都包含了一些随机性的要素。例如，行动者可能具有从一个随机分布中被赋予的初始特征。如果行动者发生互动，他们的互动模式就可能是随机选择的，等等。相同的情况在现实社会中也被假设为是真实存在的：发生什么事情有一定程度的随机性。这种随机效应就是当运行模型许多次时，每一次都将产生不同的结果(在下一节中会对此做

更详细的处理)。即使这些结果只有很小的不同,但人们希望出现的最好情况就是最频繁出现的结果——模型输出结果分布模式——与实际观察到的结果相对应(Axelrod, 1997a; Moss, 2002)。如果不能对应,人们可能就想要知道这是否因为发生在真实世界中的随机事件的特定组合是一个异常值,而且如果有可能多次"重新运行"真实世界的话,最常见的结果可能会与在模型中看到的结果更加相似!

有效性验证的技术

当验证模型的有效性时有两个领域需要被检查:首先,是在理论和理论模型之间的拟合度,第二,是在模型和模型想要模拟的真实世界现象之间的拟合度。

比较理论和模型:灵敏度分析

对理论与其模型之间拟合度的最好评估是运用理论推演出一些关于变量间期望关系形式的命题,然后检验当模型用各种不同的参数设置运行时,这些期望分布实际上是否真的会出现。每一个参数设置都对应于模型作出的一个假设。人们或者通过测量源自于经验数据的值,或者通过敏感度分析来检验每一个设置。虽然前者更可取,但是有许多参数不能被实证检验,对于这些参数而言,进行某些形式的敏感度分析是必要的。

敏感度分析的目的在于理解模型产生期望结果的条件。例如,在第 1 章中描述的民意动力学模型中,人们可能会问,极端分子必须要多极端才能使所有行动者最终都加入到一个极端的派别中去?为了找到答案,需要对一系列不确定的参数值进行模拟,参数值的范围可能是从 0.5 到 1.0 并以 0.1 为间隔(如,运行 6 次模拟)。但是,模型包含随机因素(例如,行动者"相遇"和交换意见的顺序是随机的),所以人们应该不会只满足于运行 6 次,而是会对每个不确定的参数值都运行许多次以获得一个平均值和方差。如果对每个参数设置重复运行 10 次(运行次数需要用考虑好的方差数量来选择,以便得到一个统计上有意义

的结果)，那么我们就需要执行 60 次运行。

但更糟的是，大多数模型都包含许多参数，而且它们之间的互动可能会影响模拟(如，出现极端分子群体的数量取决于极端分子的不确定性和极端分子在政治谱系中的分布，参数所具有的影响既是独立的也是组合的)，因此，在理想状态下人们需要检验全部参数的所有值在所有组合下的输出结果。即使只有一些参数，也需要天文数字般的运行次数，因此这不是一个切合实际的策略。

如果我们将每个参数的范围想象成分布在一条轴线上，所有参数的集合就界定了一个多维的参数空间，在此空间中每一个点对应一个参数值的组合。在进行一个完整的敏感度分析时，任务的参数值范围能够被量化为这个空间的体积，而且任何削减空间的方法都将减少需要的模拟运行数量。一个明显的方法是使用以前的经验知识去尽可能多地限制参数范围。例如，我们可能知道一个参数虽然理论上能够取 0 到 100 之间的任何值，但实际上它从没有被观察到有大于 10 的值。另外，我们可以通过限定我们测试的参数值范围来限制模型的适用性：我们可以说只有参数在 5 到 10 之间的某处时模型才是适用的，我们不去考察当参数超出这个值域时会发生什么。

另一种方法是对参数空间进行抽样，它能够与限定参数范围联合起来使用。这种方法并不会在空间中每一个点上执行模拟运行，而是只使用某些点。这些点可以通过随机选择，或者是有意识地选择去检视我们特别感兴趣的组合，再或者是接近于那些模拟行为的主要变化被期望发生于其间的区域选择(相变)。

这个方法的精确版本会使用一个学习算法，例如遗传算法(参见第 5 章第 2 节)，去搜索一个范围以确认某些输出结果变量取最大值或最小值的区域(Chattoe, Saam & Möhring, 2000)。

比较模型和经验数据

如在前一节中所讨论的，并不是所有的模型都被期望去匹配经验数据；可能得不到有效的数据去与那些以发展理论为目标的模型相比较，而且没有理由去进行经验测试。对于中层模型而言，检验的标准是看模拟生成的输出结果是

否在性质上与那些在社会世界中观察到的结果相似，但是并没有期望两个结果在数量上也匹配。只有被我们称之为传真模型的模型才对模拟所得数据和经验数据之间的比较有严格的要求。在这一节中我们将考虑进行这一比较的一些方法。

社会科学家习惯对从模型中得到的数据与从现实社会中收集的数据进行比较：这种比较是在一般线性回归每次计算 R^2（决定系数）时隐含进行的(Fielding & Gilbert，2000)。这种情况下的模型就是回归方程，这种方程会计算因变量的预测值。同样，人们能够测量从一个模拟模型中得到的输出变量值与经验观测值之间的拟合度(实际上就是两组值之间的相关系数)。然而，这种简单的程序却做了一些很强的假设，这些假设虽然通常能满足线性回归的要求，但却不太可能适用于模拟模型。

模拟模型的一个重要特征在于输出变量的值随着模拟的运行而变化。例如，在一个关于消费者行为的模型中，一个特定品牌的购买者数量在模拟过程中可能被观察到从 0 发展到大多数人。然后这个发展趋势可能会与真实产品的销售增长相比较。因为用于比较的数据是时间序列的，因此必须允许存在自相关的事实：在时间 $t+1$ 时的值并不独立于在时间 t 时的值。被称为 ARIMA 的统计程序能用来比较这种时间序列数据(Chatfield，2004)。

小结

在本章中，我们考虑了一些涉及设计和执行基于行动者建模研究的概念性问题。通过例子我们给出了在开始编程前需要进行的分析的类型，而且提到了在使模型具有可检验性和有效性时会出现的一些挑战。下一章我们将进入操作执行的内容：如何为模型编码，计划它的发展，以及最后给出它的结果。

第 4 章 | 设计和发展基于行动者的模型

工具箱，程序库，语言，框架和环境建模

虽然某些建模者只使用一种常见的编程语言(任何语言都能使用，但最频繁使用的是 Java)来建立他们的基于行动者模型，但这是一种较难的开始方式。多年以来，许多模型都具有相同或相似的构建模块，相互之间差异也很小。为了不继续无谓地重复相同的工作，经常被使用的元素已经被集成到可以连接进基于行动者程序的程序库或框架中。其中第一个被广泛使用的是 Swarm(http://www.swarm.org/)，虽然现在这个程序库基本上已经被完全取代了，但它的设计影响了更现代的程序库，诸如 Repast(http://repast.sourceforge.net)和 Mason(http://cs.gum.edu/~eclab/projects/mason)。后两个程序库都是用 Java 写成的，所以最容易连接进同样也是用 Java 写成的模型中，但它们也能以其他语言来使用。Repast 可以获得应用于.NET 的版本并且能够很容易地连接进用 C#，Visual Basic 和 Python 写成的程序中。两者都具有一系列相似的特征，诸如以下这些：

(1) 各种有帮助的模型范例；

(2) 为事件驱动模拟提供的一个精确的调度程序；

(3) 一些将模型以及行动者移动的空间在屏幕上可视化的工具；

(4) 将结果搜集进一个文件以用于后期统计分析的工具；

(5) 确定模型参数并在模型运行时改变它们的方法；

(6) 支持网络模型(管理行动者之间的连接)；

(7) 从模型到地理信息系统(GIS)的连接以便于环境能在真实地形上被模

型化；

(8) 一系列去除漏洞的算法，用于进化计算(第 5 章第 2 节)，随机数的产生和神经网络的执行。

许多人多年的努力都用在了这些程序库的构建上，对它们的使用极大地减少了构建模型的时间和产生错误的机会。Repast 和 Mason 都是资源开放软件，非商业性使用可以免费获取。它们仅有的不是在于：第一，复杂性，这意味着在人们能够充分利用它们具有的广泛特征之前需要花费几个月的时间；第二，建模者被期望使用一个相对低层次的语言，例如 Java，去发展他/她的模型。

更适合初学者的是建模环境，它提供了模型被创建和执行的完备系统，并且运行结果不用离开此系统就被可视化。这种环境往往更易于学习，人们在拥有一个运作的模型之前所花费的时间比人们使用程序库方法花费的时间缩短了很多。然而，简单化的代价就是灵活性更小、执行速度更慢。如果你需要它们提供更多的执行力和灵活性，就得投入时间去学习如何运用一个基于程序库的架构。

最初是为了其他目的建立的环境也能够被用于模拟，而且有时是非常有效的。例如，简单的模拟能够运用电子制表软件包 Microsoft Excel 创造出来，并且，免费的开放资源统计软件包 R(http://www.r-project.org/)对于处理大批量数据也是很有用的。一些重要的基于行动者的模型已经通过使用数学软件包 MatLab(Thorngate, 2000)和 Mathematica(Gaylord & D'Andria, 1998)创建出来。尽管如此，为基于行动者的模型而特别设计的环境通常是首选。

Repast

Repast(North, Collier & Vos, 2006)是一个有两个程序库的族。一个适用于 Java，一个适用于 Microsoft 的.NET，以及一个更可视的工具，这个工具允许使用脚本语言 Python 来发展模拟。这两个程序库在功能上是等同的，选择使用哪一个只取决于你对编程语言的选择。Python 系统更接近于模拟环境，并且不需要先进的编程技巧。

Repast 具有一个活跃的用户群——来自于学术界、政府和商业公司——和

一个通讯列表，这在你遇到关于如何使用 Repast 的问题时会非常有帮助。Java 版本几乎可以在任何平台上运行，包括 Windows，Mac OS X，和 Linux。

Mason

Mason(Luke，Cioffi-Revilla，Panait，Sullivan & Balan，2005)是另一个 Java 程序库，它也受到了 Swarm 的极大影响。它同时提供标准程序库和 3-D 可视化程序库，使用者能够在模拟运行时录制它的影像(这些特点在 NetLogo 中也可获得，参见下面的内容)。Mason 是免费的开放资源。

NetLogo

目前，最受欢迎的基于行动者的模拟环境是 NetLogo(Wilensky，1999)。它包含一个用户界面建设器以及其他的工具，例如系统动态建模器。出于教学和研究目的可以免费获取，可以从 http://ccl.northwestern.edu/netlogo/下载。它能在所有常用的操作系统中运行：Windows，Mac OS X 和 Linux。NetLogo 拥有一个活跃的用户社区，能够迅速和全面地回答使用者的问题，并且用户遍布自然科学和社会科学的所有教育层次。其他环境包括 StarLogo(http://education.mit.edu/starlogo/)和 AgentSheets(http://agentsheets.com)，但是这些环境更适合创建和展示用于教学的简单模型而不适合在研究中建立模拟。

比较

表 4.1 在一些标准上使用了公认的主观判断，比较了 Swarm，Repast，Mason 和 NetLogo 环境。不存在对于所有使用者而言都是理想的环境。为了在它们之间进行选择，人们需要考虑自己本身的专业知识和编程经验、模型可能具有的复杂性，以及研究项目的目标(如，项目是不是探索性的，模型是不是非常简单，或者项目是不是为了建立一个相对复杂的模型并尽可能地与数据进行对比以测试它的行为)。Repast、Mason 和 NetLogo 仍在继续发展并在相当迅速地进化，所以表 4.1 中的信息需要与每个系统目前的状况进行对照核实。其他的比较和回顾可以在 Castle 和 Crooks(2006)，Gilbert 和 Banks(2002)，Rails-

back,Lytinen 和 Jackson(2006),Tobias 和 Hofmann(2004)的著作中找到。

表 4.1 Swarm, Repast, Mason 和 NetLogo 的比较

	Swarm	Repast	Mason	NetLogo
许可证*	GPL	GPL	GPL	免费,但不是开放资源
文件	拼凑的,	有限的	改进中,但是有限的	好的
用户群	正在减少	大	正在增加	大
建模语言	对象的—C, Java	Java, Python	Java	NetLogo
执行速度	中等	快	最快	中等
支持图像用户界面的发展	有限的	好	好	用"点和击"很容易创建
嵌入能力以创建影像和动画	不能	能	能	能
支持系统性试验	有些可以	可以	可以	可以
学习和编程的容易度	差	中等	中等	好
安装容易度	差	中等	中等	非常容易
连接地理信息系统(GIS)	不能	能	能	不能

注:* GPL:综合公共许可证,http://www.gnu.org/copyleft/gpl.html。

NetLogo 作为最快能学会和最容易使用的程序库脱颖而出,但是它对于那些大而复杂的模型可能并不是最适用的。Mason 与 Repast 有可比性,但通常比 Repast 更快,它具有最新程序库的优势,吸收了早期系统的经验,但是只拥有一个明显小很多的用户群,这意味着只有一个较小的用户社区可以提供建议和支持。

使用 NetLogo 建立模型

在接下来的部分,我们将运用基于行动者的模拟环境 NetLogo(Wilensky, 1999)。NetLogo,如同上面提到的其他环境和程序库一样,在经历着持续的发展,每年都会出现一个主打的新版本。因此,在本册中列印出的编码(是使用版本 3.1 发展出来的)在将来的版本中运行可能就需要修改。然而,Wilensky 和他

的团队努力使 NetLogo 升级改变具有兼容性，所以当你下载编码时任何需要作出的修改可能都会自动完成，或者仅仅只需要做微小的调整。

NetLogo 系统呈现给使用者 3 个标签：界面标签，信息标签和程序标签。界面标签用于使模拟的输出结果可视化并控制它(图 4.1)。信息标签提供了关于此模拟是用于什么和什么应该被观察到这样一些基于文本的文件。程序标签是人们使用针对这个环境的特殊语言(NetLogo 语言)写的模拟程序。NetLogo 是以编程语言 Logo 为基础的(Papert，1980)。Logo 是为了教小孩关于程序和算法的概念而设计出来的，并且最初是用于控制称为“乌龟”的小玩具机器人。为了表示纪念，NetLogo 中的行动者仍然被称为“乌龟”。

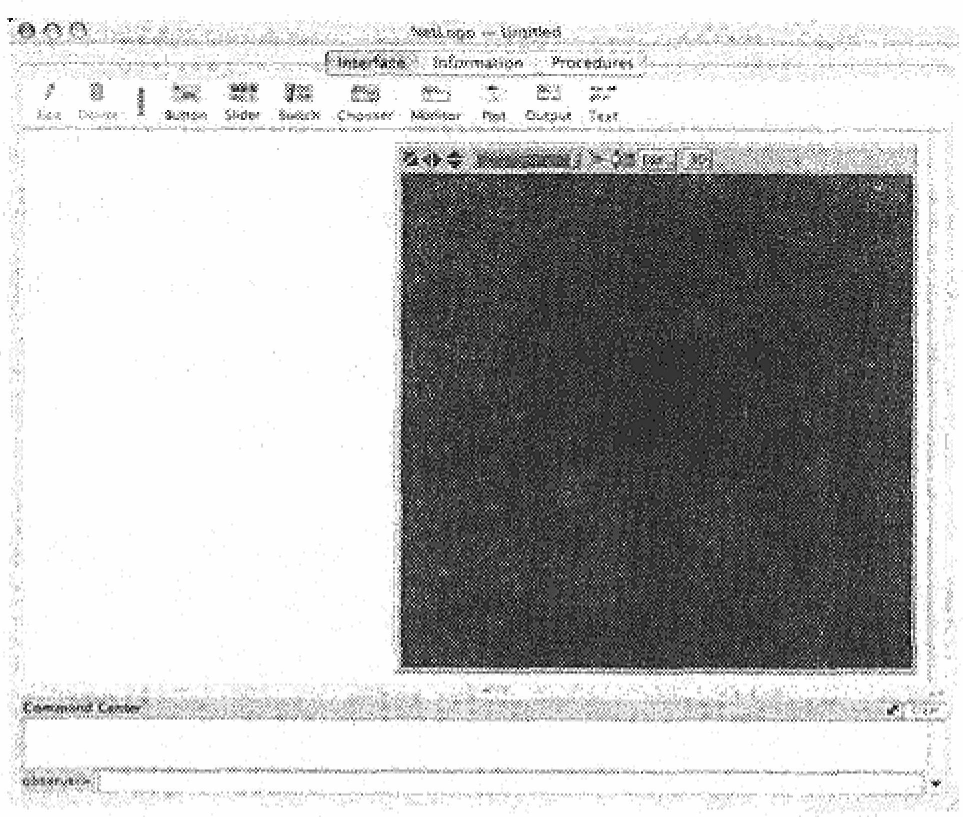


图 4.1　NetLogo 界面

界面标签里一个黑色的方格被称为视图，是由一格子的光点组成。这是行动者运动的空间环境：一个模拟程序可以指示行动者移向从光点到光点的任何方向，行动者在视图上是可见的(参见图 4.3，视图上的小三角形就是行动者)。通常，NetLogo 环境是被设置好的，以使左手边缘与右手边缘相衔接，上边缘与下边缘相衔接，所以，如果一个行动者移出了视图的左手边缘，它立即会在右手边缘重现(环境在拓扑上等同于环形面，一个圆环状的实体)。光点开始时是黑

色但能很容易地被重新着色，因此，举例而言，人们可以创建一个等高线图。视图中光点的数量也能够被设置：当 NetLogo 启动时，视图由 35×35 的光点组成，但这个数目可以增至好几千。

一个 NetLogo 程序有 3 个部分。第一，一部分说明行动者将是怎样的以及对所有行动者而言都可获得的变量名称（全局变量）。第二，具有一个安装程序来初始化模拟。第三，具有一个进行程序，它被系统重复执行以运行模拟。图 4.2 展现了一个非常简单的例子以给出一个关于 NetLogo 程序的初步印象，其中有 10 个行动者被创建出来并且无限期地随机做来回运动。

在这个程序中，没有全局变量，所以从安装程序开始。任何从先前一个运行中退出的乌龟（行动者）都会被清除出程序，10 个新的乌龟会被创造出来（这些乌龟会被放置在 NetLogo 视图的中央）。进行程序告诉每一个乌龟去执行方括号中的命令：首先，向右（如顺时针方向）转一个随机的度数，然后先前进一个单元，其中单元为一个光点边的长度。每个乌龟相对于其他乌龟都是在独立地移动，所有的移动都是同时的（因为 NetLogo 是在一个普通的电脑中运行，行动者不能都正好在相同的时间里运作，但是 NetLogo 通过使用非同步随机更新使行动者看上去好像都在同时移动，参见第 2 章第 3 节）。

```
to setup
  clear-all
  create-turtles 10
end
to go
  ask turtles [
    right(random 360)
    forward 1
    ]
end
```

图 4.2　生成 10 个行动者并使它们随机移动的 NetLogo 程序

为了能在你自己的电脑上运行这一程序，你需要下载和启动 NetLogo。然后点击程序标签并输入图 4.2 中显示的几行编码。返回到界面标签。点击位于上方的"按钮"图标，然后点击视图旁边的白色区域。NetLogo 将给出一个你应该标注为"安装"的按钮。接着完成相同的操作后会看到"进行"按钮，同时设定

复选框“永远”为打开状态(这将导致当“进行”按钮被点击时进行程序被循环执行,一直持续直到按钮被第二次点击才会停止运行)。点击你的“设置”按钮将创建 10 个乌龟,相互堆叠地显示在视图的中央。点击“进行”按钮将会指使乌龟(行动者)按照一个随机轨迹来回穿梭在视图上。再次点击“进行”按钮以停止这个程序。NetLogo 界面看起来应该类似于图 4.3。

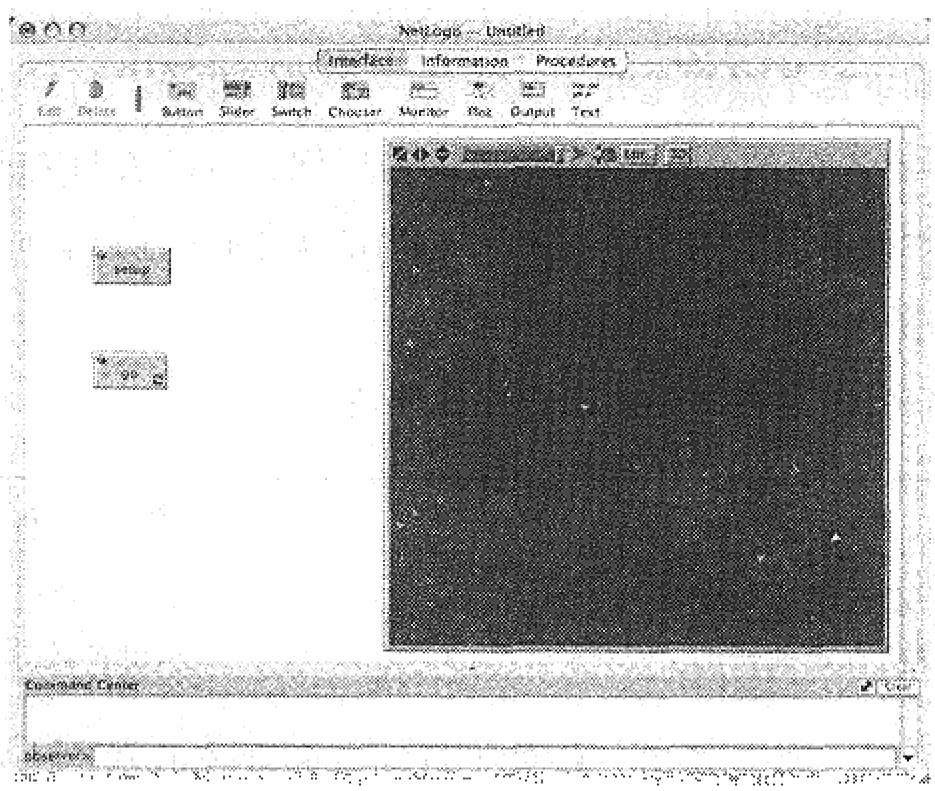


图 4.3 简单的程序运行

虽然这是一个非常简单的例子,但是它提供了一个关于人们能够多么迅速地在如 NetLogo 的环境中构建基于行动者模拟的构想。“按钮”(以及滑动条,开关等)图形可以通过“拖和放”来控制一个模拟。“视图”不需要进行任何编程就为行动者和它们环境的可视化提供了许多可能性。NetLogo 在界面标签上也能显示输出结果变量的动态变化图。虽然 NetLogo 编程语言与通常的程序语言有某些不同,但它既强大也简洁,使复杂的模拟能够用令人惊讶的寥寥几行编码就能编程出来。

本书限于篇幅,不能给 NetLogo 提供一个详细的教程,但是这个系统自身就包含了很好的嵌入式教程(可以在“帮助”菜单下找到它),而且还伴有大量的演示和模型范例,其中有一些范例还与社会科学有关。

逐步建立集合体模型

在第3章中，我们介绍了“集合体”模型。简言之，这个模型模拟了基于知识的形成物——诸如科学家社区，时尚运动和亚文化群体——的动态创建和维持。这个模型的环境是一个空间环境，代表的不是地理空间，而是一个“知识空间”，在这个空间里每个点都是知识元素的一个不同集合。行动者穿越这个空间，代表人们持有不同的(或改变了)知识和信仰。行动者只有非常简单的行为：如果他们是“孤单的”，即远离行动者的局部集中区，他们就向群体移动；如果他们是“拥挤的”，他们就离开。

因此，严格来讲，存在两种行动者行为准则：

条　件	行　动
行动者是孤单的	移向群体
行动者是拥挤的	从群体离开

建立模型的第一步是作出一些关于行动者和环境的基本决定。模型设定表明将只存在一种类型的行动者，并且行动者将在一个空间中移动。我们需要决定这个空间的维度：简化起见，我们将使用能够直接在NetLogo视图上绘制出来的一个二维网格。为了避免在网格边缘可能产生的特殊效果，我们将使用一个没有边缘的圆环面。这是NetLogo的系统默认安排，所以不需要任何额外的设定。

下一步，最好用图形或者用“伪编码”来列出模型的逻辑。为了用图形来显示逻辑，使用统一的建模语言(UML)会比较方便，它是一种展现程序的方法，这种方法是作为一种独立于编程语言细节的沟通软件方法被提出来的(Miles，2006)。UML提供了一系列标准化的图形，能够用来显示程序中对象的种类层次。这些图形包括：顺序图显示了一件事情如何引起下一件事情，与流程图相似的活动图。UML非常适合于在一篇发表的论文中描述模型，但是集合体模型太简单以至于完全不需要用到UML。例如，只有一个种类(对行动者而言)

和只有两种行动者行为(前进和转弯)。

对于这种简单的模型,另一种可供选择的方法或许更有帮助——使用"伪编码",这是一种自然语言和编程惯例非正式的混合体,它在不要求阅读者熟悉任何特定的编程语言的条件下能够使得程序的结构和流程是清晰的。图 4.4 显示了用伪编码写成的集合体程序。

Initialization

```
  Create agents and distribute them at random in
    knowledge space
```

Execution

```
  Loop forever
    Each agent:
        Counts the number of other agents within its
          local-radius
    Each agent:
        Compares the number of other agents within its
          local-radius with the threshold
        If the number is greater than the threshold
        Then(the agent is crowded)
          The agent locates that agent within
            visible-radius with the most agents
            surrounding it
          The agent moves a distance proportional to
            speed away from this agent
        Else(the agent is lonely)
          The agent locates the agent within
            visible-radius with the most agents
            surrounding it
          The agent moves a distance proportional to
            speed towards this agent
End loop
```

图 4.4　用伪编码写成的集合体程序

这个程序有两部分:初始化(在 NetLogo 中称为"安装")和执行(在 NetLogo中称为"进行")。伪编码的首行缩进有助于辨别哪行跟在哪行后面。例如,程序循环重复执行"Loop forever"和"End loop"之间的几行。模型中不变的参数用斜体字显示。

一旦人们有了程序的伪编码版本，将它转换成诸如 NetLogo 的编程语言就相对容易了，这样做的结果显示在图 4.5 中。图 4.6 是 NetLogo 环境中程序编码的截屏，图 4.7 显示了有建立模拟按钮（按此按钮执行初始化编码，从“setup”到下面 10 行的“end”语句）和运行模拟按钮（“进行”按钮，重复执行进行程序直到按钮被再次点击）的用户界面。

```
breed[agents agent]
agents-own[around visible]
to setup
  clear-all
  ask patches[set pcolor white]
  create-custom-agents 500[
    set color green
    set size 2
    ; distribute agents randomly
    setxy random-pxcor random-pycor
    set heading random 360
    ; ensure that each is on its own patch
    while [any? other-agents-here][fd 1]
    ]
end

to go
  ask agents[count-those-around]
  ask agents[move]
end

; store the number of agents surrounding me within
; local-radius units
; and the agents that I can see within visible-radius

to count-those-around
  set around count agents with[self ! = myself] in-radius
    local-radius
  set visible agents with[self ! = myself]in-radius
    visible-radius
end

to move
  if any? visible[
    ifelse around > threshold
      [; if more than threshold agents surround me,
       I'm crowded:
```

```
        ; face away from the most popular patch and
          become red
        face-away set color red]
      [; else I'm lonely: face towards the most popular spot and
        become green
        face-towards
        set color green]
    ]
  ; and move in my(new)direction
  fd random speed
  ; ensuring that I stop on an empty patch
        while[any? other-agents-here][fd1]
end
;face towards the most popular local spot
to face-towards
  face max-one-of visible[around]
end
; face away from the most popular local spot
to face-away
  set heading towards max-one-of visible[around]-180
end
```

图 4.5　集合体程序

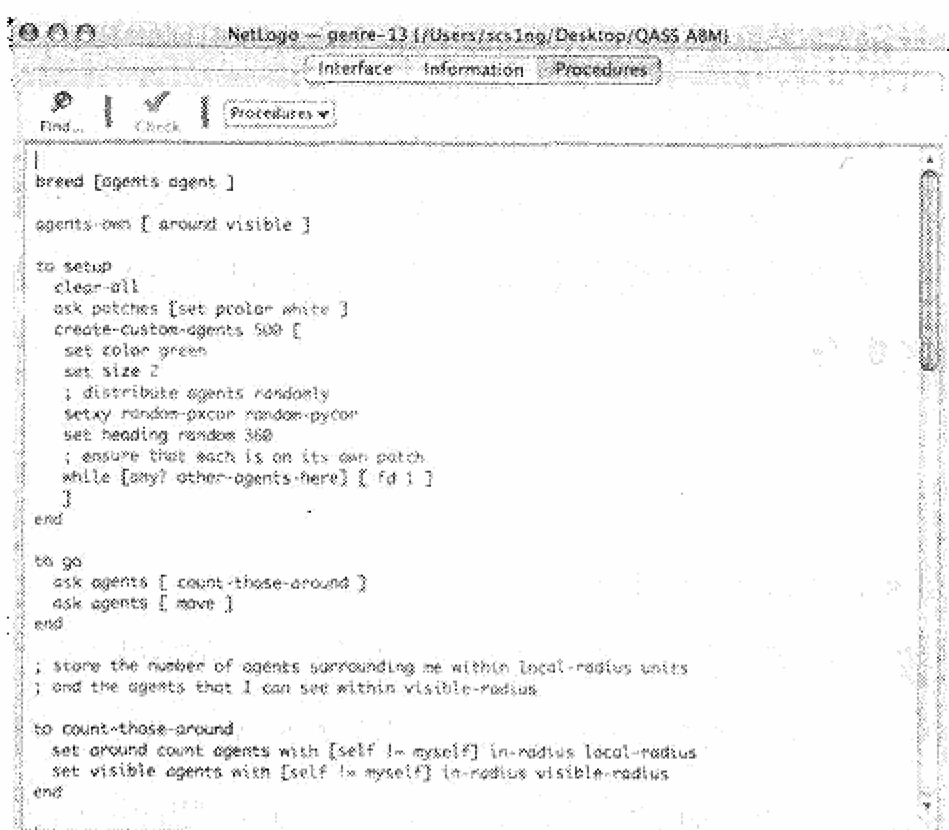


图 4.6　显示部分集合体程序的 NetLogo 程序标签

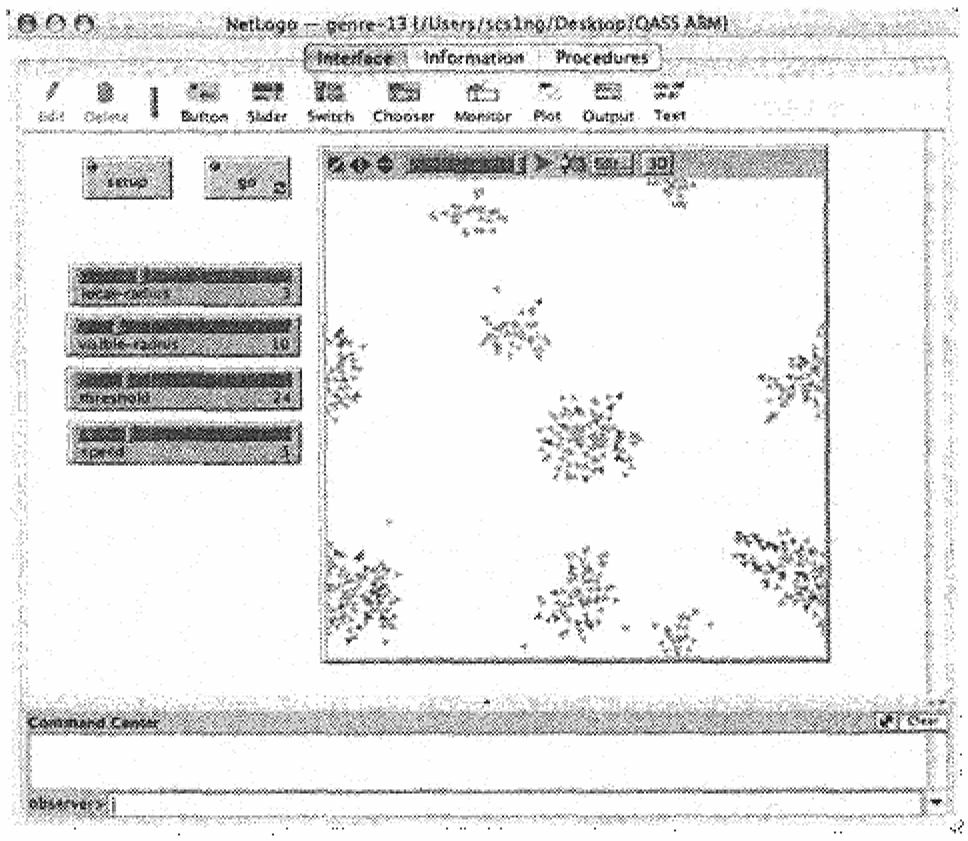


图 4.7 集合体界面

程序注释

这一节,在图 4.5 中的每一行程序都会被解释。

breed[agents agent]

这一行以复数形式和单数形式命名了将要在模型中使用的类别的名称。

Agents-own[around visible]

行动者类别有两个属性:遍及各处的和可见的。每个行动者在这两个变量上都有它自己的值,这两个变量分别储存了行动者周围其他行动者的数量和能够看到的行动者集合。

```
to setup
  clear-all
  ask patches[set pcolor white]
```

当使用者按下界面上的"安装"按钮时,安装程序就被执行。这个程序首先删除上次运行时剩下的所有行动者,然后为网格上的所有光点都着上白色。编码 ask patches[…]告诉所有光点(光点是在网格上的单元格)去执行方括号中的编码。命令项 pcolor 是一个 NetLogo 变量,用于设定光点颜色。

```
creat-custom-agents 500[
  set color green
  set size 2
```

然后500个行动者被创建。在方括号里的命令被每个新的行动者独立地执行。视图上每个行动者的图像都被着成绿色，并且其大小被设定成最小值的两倍。

```
setxy random-pxcor random-pycor
set heading random 360
```

然后每个行动者都移到网格上的随机点上(random-pxcor 产生一个对应于网格左右边缘之间一个随机位置的数字，random-pycor 在上和下边缘之间的位置上完成同样的工作)，heading(行动者面对的方向)用于设定面向一个随机的方向。

```
while[any? Other-agents-here][fd 1 ]
```

行动者随机分布的结果可能是两个或更多的行动者最后会在相同的光点上，一个行动者处于另一个行动者上面，所以这一行命令是让每个行动者去检查是否有任何其他的行动者也在同样的光点上。如果有的话，这个行动者就朝它现在面对的方向前进一个单位(fd 1)，然后再次检查，继续前进直到它找到一个未被占据的光点。

```
  ]
end
```

到此就完成了每个行动者在被创建之后就立即执行的命令，也将此 setup 程序带到了终点。

```
to go
  ask agents[count-those-around]
  ask[move]
end
```

当界面上的"Go"按钮被按下时，go 进行程序就被重复执行。每次完成一遍此程序，模拟就前进一个时间段。在每个时间段中，所有行动者都被要求去

清点网格上它们的局部区域中的行动者数量，然后所有的行动者都被要求进行移动。重要的是，在任何行动者移动之前，所有的行动者都要完成对它们周围行动者的清点工作；否则，行动者数目就可能会因为行动者在它们被清点之前就开始移动而被弄混淆。NetLogo 确保在下一个命令（在这个例子中，是第二个 ask agents[…]）开始前所有行动者都能完成构造中的命令，如 ask agents[…]。如果程序已经被写成 ask agents[count-those-around move]，就不会是这种情况了，因为每个行动者在它一完成清点任务之后就已经开始移动了，并且对于某些行动者而言，在其他行动者将它们加入计数之前这可能就已经发生了。这就是在对看起来同时行动的行动者进行编程时需要防止发生的一个关于时态问题的例子。

count-those-around 和 move 程序都是属于 agent 行动者类别的方法，即都是在程序编码中被界定去展现清点和移动对行动者而言意味着什么的程序。下面几行详细说明了 count-those-around 是由什么组成的。

```
to count-those-around
  set around count agents with[self! = myself]
    in-radius local-radius
  set visible agents with[self! = myself] in radius
    visible-radius
end
```

在这个程序的第一行里，每一个行动者都拥有变量 around，它是被设定的位于局部—半径区域半径范围内的行动者数目。使用者在模拟开始前通过移动界面上的一个滚动条来设定局部—半径的值（图 4.7）。通过从 end 开始逆向操作的方法最容易看出这行编码在做什么。编码 in-radius local-radius 生成了那些与执行程序的行动者相距大约在一个局部—半径或以内的行动者。这组行动者集包含了行动者本身（它与自己相距 0 个单位，因此比局部—半径靠得更近）。而编码 with[self!=myself]排除了行动者自身，count 反馈了剩下的行动者数量。行动者属性 around 就被设定为这个数目。人们可以想象行动者环顾四周，计算它在局部半径 local-radius 之内看到的行动者数量并将此记忆在

around 变量中。在生产系统的术语中(参见第 2 章第 1 节),around 是工作记忆的一部分。

下一行编码与前面行非常相似。它会找出能够被行动者看到的其他行动者有哪些,即哪些行动者在可见一半径之内,这是一个其值被界面上的另一个滚动条所设定的参数。第二行编码与第一行编码稍微有些不同,因为变量 visible 储存的不是可见的行动者的数量而是行动者自身的名字。这一步是必须的,以便于它们稍后能被辨认出来(在 face-towards 和 face-away 程序中)。

"移动"程序既决定行动者将要移动的方向,同时也执行这个行动。用生产系统的术语讲,移动程序是一个非常简单的规则解释器。

```
to move
    if any? Visible[
```

这个程序以通过检查对于执行这种方法的行动者而言是否存在任何可见的行动者为开始。如果存在,行动者需要决定它是"孤单的"还是"拥挤的",即它需要决定条件—行动准则是否都适用。

```
Ifelse around>threshold
  [
    Face-away
    Set color read]
```

"条件"包括检查在这个行动者周围的行动者数量是否大于用户界面上滚动条设定的门槛参数的值。如果大于的话,它就是"拥挤的",行动者就必须远离群体的中心。程序 face-away 完成这个任务。另外,为了在视图上显示这些行动者,它们的颜色被转变为红色。

```
[
    face-towards
    set color green]
]
```

另一种情况是行动者是"孤单的",它们必须移向群体并且变成绿色。

```
fd random speed
while[any? other-agent-here][fd 1]
end
```

一旦行动者已经设定好它的方向，它就能以一个被"速度"参数所决定的速率进行移动(随机速度会返回一个小于"速度"值的随机数字)。和前面一样，如果行动者坐落在一个已经被另一个行动者占据的光点上，它就继续沿着相同的方向向前移动直到找到一个空置的光点。

"移动"程序就此结束。我们仍然需要界定 face-towards 和 face-away。

```
to face-towards
    face max-one-of visible[around]
end
```

face-toward 命令检索行动者能够看到的行动者集合(已经被记忆在 visible 变量里的行动者)，并找出那个集合中围绕在每个行动者旁边的行动者有多少。旁边有最多行动者围绕的那个行动者被认为是行动者想要移向的群体的中心。NetLogo 命令 face 设定了行动者面向的那个中心行动者的方向。

```
to face-away
  set heading towards max-one-of visible[around]-180
end
```

to face-away 命令与 face-toward 命令几乎是相同的，但移动方向与最中心行动者的方向却是 180 度相反。

集合体模型的程序编码到此就结束了。

运行模型显示出行动者的初始统一随机分布分离成了"聚类"，其中一些行动者在中心，其他行动者分布在它们周围。中心行动者是拥挤的，所以会移动。它们这样做时会稍微移动聚类的中心并可能使其他行动者或者拥挤或者孤单，然后其他行动者也将移动。因此，行动者的聚类虽然会持续在一起较长时期(用时间段测量的话)，但它会漂移穿过视图。孤单的行动者会移向聚类，有时候会加入它，有时候会继续尾随其后。而聚类从来不合并。

图 4.8 展示了一个典型的模拟快照。在这个图中，拥挤的行动者是深灰色

的，而那些孤单的行动者是浅灰色的。

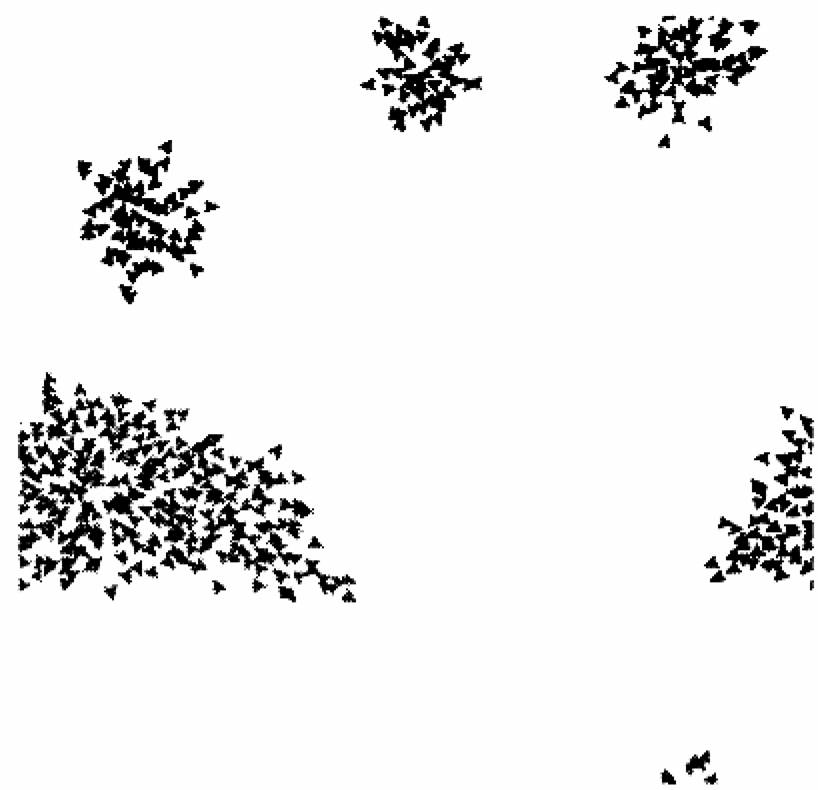


图 4.8 模拟快照

将这个模型的行为与第 3 章第 1 节中描述的集合体特征相比较，我们可以看到如下情况：

(1) 当我们运行模型时，我们看到“聚类”，但是在聚类周围划一个边界会涉及某些主观任意的界定，或许依据行动者的局部密度。

(2) 虽然关于哪些行动者在聚类内部哪些行动者在外部的界定是可能的(如根据到最近邻居的距离)，但这看上去仍然是武断的。

(3) 在相同聚类中的行动者是紧密聚合的，因此可能被认为共享了它们知识的某些方面。

(4) 聚类的方位通过它们中心的位置表现出来，但总是在不断地变化，因为一些行动者更紧密地移向聚类而另一些则寻找新的较不拥挤的地点。

(5) 某些行动者认为自己是拥挤的，并且这些行动者会和聚类中的其他行动者表现得不同(通过“创新”或通过尝试穿过知识空间来寻找较不拥挤的地方)。这些行动者在聚类中位于更中心的区域，并对设定其他行动者的移动方向具有影响力。

因此，我们在社会中观察到的集合体特征在模型中作为行动者行为的结果

出现了。虽然其他微观层次的行为能够产生相同或相似的宏观层次模式(Gilbert, 2002),但了解产生我们观察到的宏观层次行为的这些微观行动是有用的。特别需要指出的是,如果模型满足下列条件,那么我们就能得出结论说,我们描述为"集合体"的现象将从行动者的行为中产生出来:

(1) 行动者改变它们在知识空间中的观念以应对"过度拥挤"(Mulkay & Turner, 1971);

(2) 某些观念和行动者被认为是地位高的或重要的;

(3) 行动者被激励去模仿和接受这些观念或者它们的变化。

对模型输出结果和经验数据更详细的比较对于这种抽象模型而言是不适用的(参见第3章第3节)。抽象模型的价值有两层:根据 Epstein 的名言"为了解释宏观的社会模式,我们在行动者模型中产生——或'发展'——它们"(Epstein, 2007:50;也可参见 Epstein, 1999),它们能解释特殊现象的产生;而且它们有助于凸显现象之间的共同性和差异性,否则这些现象可能被认为是不可比较的。因此,这种模型成功的更好标准是它在多大程度上产生了深层的理论问题或者提供了能够从经验上确认有效性的中层理论。例如,模型提出了这一问题:假定"庞克族"和"科学家"的社会形成能够通过相同的普通模型被复制出来,那么它们之间显著的相似性和差异性是什么?在许多社会生活领域都可以发现类似的微观行为以及相应出现的许多集合体,因此这种模型可以被应用于解释广泛的社会现象。

计划一个基于行动者建模项目

像任何研究项目一样,提前就一步一步地计划好一个模拟项目是有帮助的,对你计划的可实现性也会更有信心,而且,如果你明显落后于计划的话,就能够开展补救性的工作。虽然大多数的模拟项目在本质上与使用其他研究风格的项目没有区别,但它有一些特别的特征需要注意。

(1) 不要低估它所要花费的时间。人们倾向于只考虑花费在写编码上的时间,但是,通常设计一个模型所花费的时间与写编码的时间一样长,并且经常要

花费比写程序更长的时间去清除程序漏洞。因此，如果人们估计出写一个程序所需要的时间，然后最少将它乘以3来得出构建模型所需要的总时间，这一点也不悲观。除非使用一个如NetLogo的建模环境，否则大多数的编程时间就都将被发展用户界面和结果显示程序占用，而不是被模型编码本身占用。这就是为什么建模环境如此有价值的一个原因——它们能省去大量工作。

(2) 保存日志。想法构思会在项目进行的所有阶段随时出现，你会有忘记它们的风险，除非你在日志或实验记录中将它们记下来。你要对在建立模型时遇到的问题特别注意：你最初假设只是技术性编程问题的困难可能具有广泛的重要性。例如，如果模拟的结果看上去对一个参数的特定值非常敏感，这在建立模型时可能只是一个问题，但它也可能暗示关于这个参数在真实世界中起作用的某些实质性结论。

这里有一些额外的注意事项，那些有许多研究者在一个团队中工作的大型项目尤其需要当心。

(3) 找到具有合适的技术人员。如果你是一个单独的研究者，你会知道你已经在模型编程上掌握了何种程度的技术。如果是一个更大型的项目，需要一些分工，你可能需要招募一些具有建模专业知识的人。因为基于行动者的模拟仍然是一种新方法，具有丰富经验的研究者比较难找到，你可能需要雇用一些具有其他技术的人员并对他们进行基于行动者建模的培训。熟悉在Java中编程是特别有用的技术(即便你并不打算使用一个基于Java的程序库——参见第4章第1节——Java课程给予的编程基础也是非常有用的)；在有待建模的领域有一些研究经验；以及清晰写作的能力，这在准备报告和论文时非常重要。

(4) 参加项目内部交流。如果不止一个人为项目工作，就需要注意确保每个人都相互理解并知道其他人正在做什么。虽然这在所有的团队工作中都是正确的，但在建模项目中，经常会有一些人是研究领域内的专家但对建模却知之甚少，还有另一些人是建模专家但对研究领域内的知识却知之甚少。双方可能都会对询问问题以及向其他人暴露自己的无知感到不自在。在大项目中，可能有必要安排特别的培训环节，其中在项目特定方面具有更多知识的那些人可以帮助其他人，以便将每个人都带到一个共同的知识和技术水平上。

(5) 注意时间安排。大多数的建模项目都会设计一些数据收集和模型发展的工作。如果模型设定要等待经验数据的收集和分析，这在时间安排上可能比较棘手，但是数据的收集依赖于事先对测量对象的精确界定。除非保持注意，否则就会陷入第 22 条军规的困境，在这种情形下建模者和数据收集者都不能有所行动。

报告基于行动者的模型研究

学习如何报告基于行动者建模结果最好的方法是学习其他人已经做过的报告。以已经找到的有帮助作用或有趣的论文为例，仔细地去看作者是如何构建它们以及什么使它们具有说服力。虽然基于行动者建模是一个比较崭新的领域，还没能发展出一套关于如何写作论文的健全的惯例规范，但是也有一些共同的要素(Axelrod, 1997a)。在 Richiardi, Leombruni, Saam 和 Sonnessa (2006)的著作中可以找到一些有帮助性的论述。

一篇基于行动者建模报告或期刊文章的主要部分通常如下：

(1) 摘要，这应该指出(大致以此顺序)

a. 论文中思考的主要研究问题；

b. 论文的发现和结论；

c. 使用的方法(例如：基于行动者建模，抽样调查分析)，以及对经验数据而言从中收集数据的样本。

(2) 序言，介绍论文中要解决问题的背景并解释这个问题为什么是令人感兴趣的。

(3) 文献综述，论述以前的研究，并指明为什么这篇论文报告的研究对以前的研究是有价值的补充或改进。关于研究问题或领域的文献，还有关于相关模型的文献，甚至那些先前还没有应用到这个领域的模型都应该回顾。这一部分应该明确指出报告的研究中哪方面是一个进步以及它如何借鉴了以前的研究。

(4) 对你想要解释的规律的陈述(这通常是对导言和综述内容的一个总结)。它们可以作为一组你想要证实(或证伪)的正式的假设被提出来，或者也

可以通过不那么正式的形式呈现出来。

(5) 对模型的描述。这个描述需要足够详尽，在原则上，它应该能使读者可以重新执行你的模型并得到相同的结果，但它不用包含程序编码(某些读者不会知道和理解你使用的编程语言)。作为替代，可以使用图表(如 UML)或伪编码来描述你的模型(参见第 4 章第 3 节)。特别要注意模型中事件发生的顺序：这在重新准确地运行模型时最容易导致问题出现。不要担心将与变量相关的公式包含进来，如果这些公式对精确地设定你的模型有帮助的话。

(6) 参数描述。你为每个参数所选择的值都需要解释和合理化。一些值的选择可能是基于对社会世界的观察(如：在劳动力市场模型中的就业率)；一些值的选择可能是合理的猜测，而且你已经用敏感度分析考察了它们的值变化时的影响；一些值的选择可能被用做“反向”推论，因为只有这些值给出了你想要用模拟去演示的输出结果模式。所有这些都需要被解释。

(7) 结果描述。这几乎肯定涉及对图表的呈现和评论，这些图表显示了你从模拟运行中观察到的变量是如何相关的。要认真去弄清楚执行这些模拟的条件。例如，这些图是否显示了多次模拟运行的平均值，如果显示了的话，运行了多少次以及运行之间的差异有多大(你可能考虑使用误差条线去显示差异程度)？如果你给出一个变量从零时间段随时间推移向前发展的趋势，那么就要确保你绘制的运行图足够长，从而能够清楚地表明这个趋势已经变得稳定并且不会正好在图形之外发生剧烈的变化(Galan & Izquierdo, 2005)。如果你在一个特定的时间段上将变量的值关联起来，就要确保你说明了测量进行的时间段。

(8) 采取了哪些模型检验(参见第 3 章第 2 节)和有效性确认(参见第 3 章第 3 节)步骤的讨论，以及读者对你的结果有多大信心的讨论。

(9) 结论。这里应该提出列在(4)中的假设，并清楚地说明模型是否证明了它们是正确的，还是错误的，或者是还未能被证明。然后这一部分可以对序言中的构想做进一步的发展，提出一个一般性的结论，可能的话对其中的意义进行一些思索(例如，如果这篇论文是关于劳动力市场的，就要考虑有什么样的政策可能在降低失业率上取得进展)。

(10) 致谢。简要地感谢赞助者,基金提供者以及那些帮助你进行研究的人们。

(11) 参考文献。只需要包含论文中引用的那些著作而不用列出其他。一般而言,对于你希望能够发表的期刊所要求的格式,你要确保提供完整的参考书目细节。

(12) 附录(可供选择)。可以放置较大的表格或者可能是你模型的伪编码版本。

小结

这一章描述了执行一个基于行动者的模型的全过程,从选择一个工具箱到报告结果。最重要的步骤是在你完全投入这个工作之前就先计划你的工作。这对确保你有一个明确界定的要回答的研究问题,以及允许你有足够的时间和资源去回答它都将会有很大的帮助。

第 5 章 | 基于行动者建模的发展

基于行动者建模是一个快速发展的领域，新的方法在被不断地介绍。在这一章中，我们将简要介绍目前研究中的一些新的发展方向。

地理信息系统

大多数基于行动者模型都会有行动者在一块地域上移动，但通常这块地域是一个矩形平面或一个圆环面。在这些抽象的平面空间中，比较新的模拟环境为创建复杂的虚拟平面提供了可能性，或者引入了依据真实地貌绘制的地形图。它们是通过将一个地理信息系统(GIS)集成进模型来做到这一点的。GIS是一个管理、储存和显示空间数据(如地图)的软件系统(Chang, 2004; Heywood, Cornelius & Carver, 2002)。GIS 研究本身就是一个专业，并且这也是为什么将 GIS 运用于基于行动者建模发展缓慢的原因之一：GIS 有它们自己的软件技术和技巧，这些技术和技巧与基于行动者建模工具的融合存在一些困难。

GIS 将空间数据存储在特殊建构或改编的“有空间意识的”数据库系统中。这些数据库系统是被设计用来有效地回答人们在管理地理数据时的各种问题，诸如“告诉我离这个对象不超过 10 个单位远的其他所有对象”。GIS 数据一般被安排在图层上，图层将数据包含在一个或一些变量里。当展现或操作一幅地图时，人们可以打开或关闭一些图层，选择只在可见的图层里看到变量。例如，一幅地图可以分别在不同图层里包含道路和湖泊。如果有人想看公路网络，但对湖泊不感兴趣，湖泊的图层就可以关闭。GIS 有两种类型：光栅和向量。光栅

是规则的单元格网格，每个单元格储存一个值。光栅变量适用于在平面上连续变化的数据，比如日照的高度和时间。向量变量适用于以点，线或面为形式的数据(在 GIS 里被称为多边形)。空间数据可以用投影图来呈现，投影图是显示一个三维体表面的方法，诸如在一幅二维地图上的地球。存在众多不同种类的投射法可供选择，虽然对于应该使用哪种投射法这样的问题可能更困扰那些想要绘制整个地球的地理学家而不是只想绘制一个镇或地区的建模者，但是这些镇或地区已经足够小以至于在可选择的投射法之间的差异已经变得无关紧要了。

使用 GIS 建立一个行动者在较真实的地域上穿行(例如只沿着城市街道移动)的基于行动者的模型是有可能的。这对于诸如为交通流量(Raney, Cetin, Vollmy & Nagel, 2002)和流行病传播(Dibble & Feldman, 2004; Dunham, 2005)建模的项目而言是重要的，虽然这可能会分散注意力，因为人们的目的可能是为一个并不在任何特定地点的抽象过程建模。需要较早作出决定的是在整个模拟运行过程中是否将 GIS 地形的不变作为充分条件，或者说，地形是否需要剧烈变动。比如，静态地呈现一个城市对交通模型而言可能就足够了，因为在模拟呈现的时间段里(几小时或几天)，城市街道的布局不会发生实质性的变化。另一方面，关于飓风对城市人口影响的模型可能需要在模拟过程中对模型地形进行更新，以考虑洪水、封闭的街道和暴风雨的破坏。然而，管理随时间变化的数据是 GIS 目前能达到的技术水平，但如果人们想要把它整合进基于行动者建模将会是一项挑战。

在设计有空间意识的行动者时，需要考虑的第二个问题是这些行动者如何探测他们所穿越地形的特征的。例如，如果要保证行人沿着街道行走，并且不会穿过建筑物，对行动者而言就需要有一种方法用以决定前面的路是一条人行道。最好的解决办法是向 GIS 发送一个查询以决定行动者前方物体的位置，然后分解回复的答案以检验向前运动不会受到阻碍。原则上，这并不复杂，但在实践中，它可能被后台运行的 GIS 和行动者占用大量的进程空间并减慢模型运行的速度。这些都是当前这代模型构建者们所面临的问题，但在接下来的几年里，它们将会被解决，并且有空间意识的基于行动者的建模架构将变得更加普及并更容易使用。

学习

到目前为止，已经提到的大多数模型中的行动者都不能从经验中进行学习。受生产准则系统驱动的简单行动者具有记录他们当前和过去状态的记忆，所以当他们经过环境时会“学习”关于环境的状态。但通常我们所说的学习意味着比这更多的东西。一种是学习更有效率的准则，然后以改变行动作为学习的结果。在先前的一节中，我们简单接触了一种允许进行这种学习的技术；虽然神经网络(参见第 2 章第 1 节)并不是基于明确的准则，但它们进行的学习与准则学习相类似。此外，还有另外两种模型化行动者常用的学习技术：强化学习(RL)和进化计算。两者都是从人工智能中借用过来的，在人工智能中学习方法的发展是机器学习，这标志着下一个重要的主题。

强化学习

强化学习(RL)类似于当我们通过反复试错的方式探索一个新城市以发现旅游胜地时所做的事情：你尝试一条街道，如果它看起来前景不错你就继续，如果没有希望，你循原路返回再尝试另一条。进行强化学习的行动者具有一种状态(如他们目前的位置)和一组可能的行动集合(如向北、东、南和西移动一个单元格)。环境被创建起来，它最终也将作出回报(到达名胜时的喜悦)。行动者必须找到一个能最大化他所获取的回报的策略(Sutton & Barto, 1998)。通常，在通往目标的道路上每多出一步都会减少最终的回报，所以目标并不仅仅是发现回报，而是要用最有效率的方式去实现它。行动者能够尝试很多次去找到最好的策略，探索不同的路径，所以，设计强化学习算法中的一个问题就是在探索(尝试新的路径，但可能是不太有效率的和浪费时间的)和开发(保持旧的途径，但可能并不是最好的方法)之间的平衡。强化学习已经被用于模型化公司学习其顾客的偏好(Sallans, Pfister, Karatzoglou & Dorffoner, 2003)和组织的决策制定(Sun & Naveh, 2004)，以及其他现象上。一些作者已经用强化学习的简单模型去解释和预测在人类主体的实验性博弈理论情境下的行为(Chen &

Tang，1998；Erev & Roth，1998；Flache & Macy，2002；Macy & Flache，2002；Mookherjee & Sopher，1994，1997）。

强化学习很适用于环境和报酬结构在模拟运行期间保持不变的问题。另一方面，如果环境是动态的，强化学习就会表现不佳。特别需要指出的是，如果所有行动者都在学习，而且一个行动者的行动会影响到其他行动者的状态和回报，那么使用强化学习就可能不是一个好的选择。

进化计算

与学习非常不同的方法是进化计算，这组技术中最简单和最为人所知的是遗传算法（其他的有进化编程、进化策略、遗传编程和学习分类器系统）。进化算法（Eiben & Smith，2003；Engelbrecht，2002）基于自然选择并涉及两个基本过程：选择和繁殖。进化计算的基础是一个群体，它的成员通过繁殖形成个体的连续世代，这些个体继承了它们父母的特征。成功繁殖的概率是由个体的适应性决定的：适应的个体比不适应的个体更有可能进行繁殖并将它们的特征传递给它们的子代。

进入遗传算法（Holland，1975）的个体有可能，但并不需要，成为模拟中的行动者。例如，在一个关于产业部门的模型中，行动者是公司，这个产业部门作为一个整体，通过公司的倒闭和成立的世代更替学习如何在产业区域中变得有生产力（Marengo，1992）。另一方面，每个行动者都可能学习使用作用于行动者准则的遗传算法，一旦考虑遗传算法，每个准则就都成为一个“个体”（这就是学习分类器系统运作的方法）（Bull，2004）。股票市场交易的模拟就是用这些行动者建立起来的（Johnson，2002；LeBaron，Arthur & Palmer，1999）。

对于遗传算法而言，每个个体都必须有它自己编码于染色体中的特征。编码可能是二进制字符串的形式，0 代表特征的缺失，1 代表存在，或者代表一个更复杂的结构。同样重要的是，对每个个体的适应性进行评估是可能的。例如，适应性可以通过公司行动者的积累资本存量测量出来，或者通过利他主义模型中的合作策略的有效性测量出来。个体依据它们适应性的比例被遗传算法选择出来进行繁殖，所以具有很强适应性的个体就非常有可能进行繁殖，而

非常不适应的个体则不可能进行繁殖。繁殖通过运用一个被称为“交叉”的过程将父母的染色体进行结合来完成。染色体切片从每条染色体中提取出来并结合成一个新的染色体，它的一部分来自于父亲或母亲一方的染色体，还有一部分来自于其他行动者。另外，为了确保即使在进行了许多混杂繁殖之后，群体中仍然能够继续存在某些差异，子代的染色体中的一些片断会被随机改变或变异。

子代的适应性会被评估以决定他繁殖孙代的可能性。最终，那些相对而言不适应的个体逐渐灭绝，并被那些由于继承了父母的适应性而更可能使他们自身具有适应性的个体所取代。虽然并没有个体进行任何学习来作为遗传算法的结果，但是群体作为一个整体却被认为在进行学习或最优化，因为其中的个体变得更适应了。

遗传算法是一个非常有效的最优化工具，但它是否是社会现象的好模型则有更大的争议性(Chattoe, 1998)。一个难点在于通常很难知道测量“适应性”的合适方法是什么，甚或很难知道当把这个概念应用于社会现象时它是否有任何清晰的意义。有时候，这并不重要。比如，在一个简单社会的模型中，人们并不需要对适应性有一个仔细推敲的定义：如果行动者有机会并且在这之前没有因为缺乏资源而死去，它们就能被设计出来进行繁殖。基本遗传算法的另一个普遍问题在于，将一个行动者所有显著的特征都编码成染色体的二进制字符串是困难的。然而，其他进化算法技术允许更多种类的结构。例如，遗传编程(Banzhaf, 1998; Koza, 1992, 1994)用程序编码来替代二进制字符串；学习分类器系统(Bill, 2004; Meyer & Hufschlag, 2006)使用条件—行动准则作为染色体的等价物。

模拟语言

多行动者模型一个最显著的特征是，行动者有互动的潜力。正是这一点将多行动者模型与微观模拟或基于公式模型(参见第1章第4节)区分开来。互动可能表现为对其他行动者存在的一个简单感知，去避开他们或模仿他们，或者，

互动也可能是涉及知识、观点或信仰的更复杂交流，这取决于被模拟领域的要求。然而，即使是在最精致的模型中，通常的情况也是行动者的互动是通过行动者对行动者无中介的、直接的信息传递而产生的。这与人类的交流有很大差别，在人类交流中，我们的思想（和意图）必须通过一个外在的自然语言来表达，这种自然语言必然是模糊不清的，它的意义必须通过学习获得而不是与生俱来的。

除了不真实之外，从行动者到行动者直接信息传递的模型会对更广范围的社会现象留下一个误导性的印象，从合作到领导，当通过使用从一个行动者准则系统直接传到另一个系统的无误的和清楚的信息进行模型化时，所有这些社会现象都改变了它们的特征。为了部分地回应这个问题，并且部分是因为在语言学中语言如何演化是一个重要的课题，用此关于“语言进化”的研究在过去的十年里在数量上开始逐渐增多并精致化（Perfors，2002）。这种工作许多都是专业化的、复杂的，在此我们对该话题只做一个相当简单的介绍并提供一些关于语言如何能够进化的线索。

假设我们有两个行动者，Adam 和 Betty，他们在一个虚拟的地形中站在彼此的附近。两者都能看到一个红色的小球和一个蓝色的大立方体，以及附近其他一些对象。Adam 想通过话语（通过从一个行动者向另一个行动者传送一串字符来执行）将 Betty 的注意力引向红球。最初，行动者没有一种共享语言用以谈论他们能够看到什么。虽然 Adam 能够表达音节的某种次序，但是 Betty 根本不能理解这一次序的意义，并且没有办法知道 Adam 正在提到球而不是正方体抑或是他们都能够看见的其他物体。问题甚至比这更复杂，因为刚开始时 Adam 并没有“球”或“立方体”的概念，两个行动者开始时具备的只有探测诸如颜色、形状、大小和方位等特征的能力。其余的一切都必须学习，目标就是进化出共享的词汇，就是一个“对一组差异的共识”（Hutchins & Hazlehurst，1995：151）。

下面是 Adam 和 Betty 可以使用的一个算法（Vogt，2005；Vogt & Coumans，2003）。Adam 开始时看着球，并看到它是圆的、红的和小的。这些特征都将球与方的、蓝的和大的立方体区分开来。Adam 能够将它归类为一个“圆

的、红的、小的"物体,并且 Adam 从此以后就能将这些球与方的、蓝的、大的物体区分开来。如果在场景中有两个圆的、红的、小的物体,Adam 就需要再找出其他特征以将球与另一个物体区分开来,比如说物体与 Adam 的距离。对 Adam 而言,"红色"可能就是一系列色度:具有不同的类别因而针对每种红色色度都有一个不同的名字,但这没有多大的帮助,所以颜色空间就被行动者的感知系统粗糙地割裂成许多小段,并且相同的过程也被应用到对其他特征诸如形状和距离的感知上。

通过这种方法,Adam 能够以他的视角来区分物体,赋予每个物体一组范畴,这组范畴是被选出来用于使物体能够与以前遇到的或目前在场的其他物体相区别。这组范畴被称为概念。所有这些都是在行动者的"头脑"中发生的。Betty 也将看到这个场景并且也会为球提炼出一个概念,这个概念很可能与 Adam 的概念相似(但是因为他们站在不同的位置上,所以对球拥有不同的视角,Betty 的概念就可能与 Adam 的概念不完全相同)。Adam 现在必须用一句话把"球"的信号传送给 Betty。假设这是 Adam 第一次谈论球,他就要为他的概念发明一个词——任何字符串都行,只要不同于他所知道的任何其他词。Adam"表达出"这个词并与 Betty 交流。然而,Betty 在这个阶段还不明白这个词的意义。

为了学习一个共同的词汇,Adam 和 Betty 参与了一个语言游戏(Steels, 2004)。这涉及重复进行的对话,其中有一个分值测量 Betty 所有概念之间的相关性,而且,我们正在讨论的词汇依据她是否断定自己已经成功地理解了话语而被更新。这个断定能够基于某些非言语的强化(如,Adam 指向物体,Betty 基于对单词的理解或者 Adam 对 Betty 成功完成一个行动的奖励)或是在其他情景中对此单词的相似使用,虽然后者需要做更多的推论并更容易出错。当 Betty 选择了一个概念与此单词相联系(或更准确地说,是将她知道的每个概念与被这个单词表达的可能性相联系),她就扮演了讲话者的角色。如果她想要表达"圆的、红的、小的"球时,她将选择与那个概念联系最紧密的单词,即她现在已经从 Adam 那里学会的单词。

这些技术(诸如我们已经勾画出的一种技术)将允许行动者去发展一种共

享词库用来描述他们都能看到的场景。但这仅仅只是一个开始，为了进行有效的交流，行动者需要一种共享的语法以及句法。他们不仅能够把使用语言作为传递说明的一种方法，同时也能把使用语言作为执行诸如命令、询问、说服等行动的一种方式。我们仅仅只是探索如何构建行动者能够在其中发展出此类能力的系统(Cangelosi & Parisi, 2002)，但在今后 10 多年里如果在这方面没有巨大进步的话，将会让人非常惊讶。

资　源

团体与协会

有 3 个区域性协会推进社会模拟和基于行动者建模的发展，每个协会都举办年会。每隔一年，他们共同组织一次世界大会。

- 北美计算机社会和组织科学协会（North American Association for Computational Social and Organization Science [NAACSOS]）。网址：http://www.casos.cs.cmu.edu/naacsos/
- 社会系统科学中基于行动者方法的环太平洋亚洲协会（Pacific Asian Association for Agent-Based Approach in Social Systems Sciences [PAAA]）。网址：http://www.paaa.econ.kyoto-u.ac.jp/
- 欧洲社会模拟协会（European Social Simulation Association [ESSA]）。网址：http://essa.eu.org/

你可以缴纳适量的年费，加入这些协会，他们对进入基于行动者建模研究社区提供了非常有用的帮助。

期刊

使用基于行动者建模的研究既出现在特定的学科期刊中，也出现在关注社会模拟的跨学科期刊中。2 个最著名的跨学科期刊为：

- 虚拟社会和社会模拟期刊(*Journal of Artificial Societies and Social Simulation*[*JASSS*])。这是一个在线电子期刊,在互联网上就可以阅读。它是免费提供的。在 http://jasss.soc.surrey.ac.uk/JASSS.html 这个网站的首页上,你只需注册便可以在每一期发行时收到一封电子邮件(每年 4 期)。
- 计算机和数学组织理论(*Computational and Mathematical Organization Theory*[*CMOT*])。这是一本印刷的纸质期刊,但也有一个在线收费的版本:http://springerlink.metapress.com/content/1572-9346/。

其他偶尔发表基于行动者建模论文的期刊包括:

- *Artificial Life*
- *Complexity*
- *Computational Economics*
- *Ecology and Society*
- *Emergence: Complexity and Organization*
- *Environment and Planning B*
- *Environmental Modeling and Software*
- *Journal of Economic Dynamics and Control*
- *Physica A: Statistical and Theoretical Physics*
- *Simulation and Gaming*
- *Simulation Modeling Practice and Theory*
- *Social Networks*
- *Social Science Computer Review*

通讯名单和网址

SIMSOC 的电子邮件发送名单会定期发布即将召开的会议,工作坊和基于行动者建模者可能会感兴趣的工作。你可以在 http://jiscmail.ac.uk/lists/

simsoc. html 这个网站加入这个名单或查看名单文件。

有一个由 Leigh Tesfatsion 维护的被称为“基于行动者的计算机经济学”的精彩网站。上面有许多精心分类的、能够连接到一系列范围广泛的文献的链接。网址：http://www. econ. iastate. edu/tesfatsi/ace. htm.

本书的网址为 http://cress. soc. surrey. ac. uk/qasss/。在这里你能找到集合体模型的编码和连接到基于行动者建模资源的链接。

注释

[1] 尽管如此，在许多情境下，诸如在模型化股票市场时，我们已经发现，缺乏理性（或者“0 智商”）的行动者的集合行为与在宏观层次上观察到的行为惊人地匹配（Chan, LeBaron, Lo & Poggio, 1999; Farmer, Patelli & Zovko, 2005）。

[2] 许多不同作者的后续论文已经使用这些标签来探讨行动者群体之间差异的出现（如 Edmonds, 2006; Hales, 2000, 2002; Riolo, Cohen & Axelrod, 2001）。

参考文献

Acerbi, A., & Parisi, D. (2006). Cultural transmission between and within generations. *Journal of Artificial Societies and Social Simulation*, *9*(1). Retrieved from http://jasss.soc.surrey.ac.uk/9/1/9.html.

Ahrweiler, P., & Gilbert, N. (Eds.). (1998). *Computer simulation in science and technology studies*. Berlin: Springer.

Ahrweiler, P., Pyka, A., & Gilbert, N. (2004). Simulating Knowledge Dynamic in Innovation Network(SKIN). In R. Leombruni & M. Richiardi(Eds.) *Industry and labor dynamics: The agent-based computational economics approach*. Singapore: World Scientific Press.

Albino, V., Carbonara, N., & Giannoccaro, I. (2003). Coordination mechanisms based on cooperation and competition within industrial districts: An agent-based computational approach. *Journal of Artificial Societies and Social Simulation*, *6*(4). Retrieved from http://jasss.soc.surrey.ac.uk/6/4/3.html.

Allen, P. M. (1997). *Cities and regions as self-organizing systems*. Amsterdam: Gordon and Breach.

Axelrod, R. (1997a). Advancing the art of simulation in the social sciences. In R. Conte, R. Hegselmann, & P. Terna(Eds.) *Simulating social phenomena*(pp. 21—40). Berlin: Springer.

Axelrod, R. M. (1997b). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.

Axelrod, R. (1997c). The dissemination of culture. *Journal of Conflict Resolution*. *41*(2), 203—226.

Axelrod, R. M. & Dawkins, R. (1990). *The evolution of cooperation*. Harmondsworth, UK: Penguin.

Bagnall, A. J. & Smith, G. D. (2005). A multiagent model of the UK market in electricity generation. *Evolutionary Computation*, *9*(5), 522—536.

Banzhaf, W. (1998). *Genetic programming: An introduction: On the automatic evolution of computer programs and its applications*. San Francisco: Morgan Kaufmann.

Barabási, A.-L. (2003). *Linked: How everything is connected to everything else and what it means for business, science, and every day life*. New York: Plume.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509—512.

Barreteau, O. (2003). Our companion modeling approach. *Journal of Artificial Societies and Social Simulation*, *6*(2). Retrieved from http://jasss.soc.surrey.ac.uk/6/2/1.html.

Barreteau, O., Bousquet, F., & Attonaty, J.-M. (2001). Role-playing games for opening the black box of multi-agent systems: Method and lessons of its application to Senegal River Valley irrigated systems. *Journal of Artificial Societies and Social Simulation*, *4*(2). Retrieved from http://jasss.soc.surrey.ac.uk/4/2/5.html.

Barreteau, O., Le Page, C., & D'Aquino, P. (2003). Role-playing games, models and negotiation

processes. *Journal of Artificial Societies and Social Simulation*, *6* (2). Retrieved from http://jasss.soc.surrey.ac.uk/6/2/10.html.

Batten, D., & Grozev, G. (2006). NEMSIM: Finding ways to reduce greenhouse gas emissions using multi-agent electricity modelling. In P. Perez & D. Batten (Eds.), *Complex science for a complex world* (pp. 227—252). Canberra: Australian National University.

Beltratti, A., Margarita, S., & Terrna, P. (1996). *Neural networks for economic and financial modeling*. London: International Thomson Computer Press.

Benenson, I., Omer, I., & Hatna, E. (2002). Entity-based modeling of urban residential dynamics: The case of Yaffo, Tel-Aviv. *Environment and Planning B*, *29*, 491—512.

Boden, M. A. (1988). *Computer models of mind: Computational approaches in theoretical psychology*. Cambridge, UK: Cambridge University Press.

Boero, R., Castellani, M., & Squazzoni, F. (2004). Micro behavioural attitude and macro technological adaptation in industrial districts: An agent-based prototype. *Journal of Artificial Societies and Social Simulation*, *7* (2). Retrieved from http://jasss.soc.surrey.ac.uk/7/2/1.html.

Boero, R., & Squazzoni, F. (2005). Does empirical embeddedness matter? Methodological issues on agent based models for analytical social science. *Journal of Artificial Societies and Social Simulation*, *8* (4). Retrieved from http://jasss.soc.surrey.ac.uk/8/4/6.html.

Borrelli, F., Ponsiglione, C., Iandoli, L., & Zollo, G. (2005). Inter-organizational learning and collective memory in small firms clusters: An agent-based approach. *Journal of Artificial Societies and Social Simulation*, *8* (3). Retrieved from http://jasss.soc.surrey.ac.uk/8/3/4.html.

Bourdieu, P. (1986). *Distinction: A social critique of the judgment of taste* (R. Nice, Trans.). New York: Routledge.

Brenner, T. (2001). Simulating the evolution of localized industrial clusters: A identification of the basic mechanisms. *Journal of Artificial Societies and Social Simulation*, *4* (3). http://jasss.soc.surrey.ac.uk/4/3/4.html.

Brown, L., & Harding, A. (2002). Social modeling and public policy: Application of micro-simulation modeling in Australia. *Journal of Artificial Societies and Social Simulation*, *5* (4). Retrieved from http://jasss.soc.surrey.ac.uk/5/4/6.html.

Bruch, E. E., & Mare, R. D. (2006). Neighborhood choice and neighborhood change. *American Journal of Sociology*, *112* (4), 667—709.

Bull, L. (2004). Learning classifier systems: A brief introduction. In L. Bull (Ed.), *Applications of learning classifier systems* (pp. 3—14). New York: Springer.

Bunn, D. W., & Oliveira, F. S. (2001). Agent-based simulation: an application to the new electricity trading arrangement of England and Wales. *Evolutionary Computation*, *5* (5), 493—503.

Cangelosi, A., & Parisi, D. (2002). *Simulating the evolution of language*. London: Springer.

Carley, K. M. (1999). On generating hypotheses using computer simulations. *Systems Engineering*, *2* (2), 69—77.

Castle, C. J. E., & Crooks, A. T. (2006). *Principles and concepts of agent-based modeling for*

developing geospatial simulations. University College London. Retrieved from http://www.casa,ucl.ac.uk/publications/workingPaperDetail.asp? ID=110.

Cecconi, F., & Parisi, D. (1998). Individual versus social survival strategies. *Journal of Artificial Societies and Social Simulation*, *1*(2). Retrieved from http://jasss.soc.surrey.ac.uk/1/2/1.html.

Cederman, L.-E. (1997). *Emergent actors in world politics: How states and nations develop and dissolve*. Princeton, NJ: Princeton University Press.

Chan, N., LeBaron, B., Lo, A., & Poggio, T. (1999). Agent-based models of financial markets: A comparison with experimental markets. *MIT Artificial Markets Project*, *Paper No. 124*, *September*. Retrieved from http://ebusiness.mit.edu/research/papers/124% 20% Poggio% 20Lo,%20Agent-based%20Models.pdf.

Chang, K.-T. (2004). *Introduction to geographic information systems* (2nd ed.). Boston: McGraw-Hill.

Chatfield, C. (2004). *The analysis of time series: An introduction* (6th ed.) Boca Raton, FL: Chapman & Hall/CRC.

Chattoe, E. (1998). Just how(un)realistic are evolutionary algorithms as representations of social processes? *Journal of Artificial Societies and Social Simulation*, *1* (3). Retrieved from http://jasss.soc.surrey.ac.uk/1/3/2.html.

Chattoe, E., Saam, N. J., & Möhring, M. (2000). Sensitivity analysis in the social sciences: Problems and prospects. In K. Troitzsch, N. Gilbert, & R. Suleiman(Eds.), *Tools and techniques for social science simulation*. New York: Physica-Verlag.

Chen, Y., & Tang, F. (1998). Learning and incentive-compatible mechanisms for public goods provision: An experimental study. *Journal of Political Economy*, *106*, 633—662.

Clark, W. A. (1991). Residential preferences and neighborhood racial segregation: A test of the Schelling segregation model. *Demography*, *28*(1), 1—19.

Cobb, C. W., & Douglas, P. H. (1928). A theory of production. *American Economic Review*, *18* (Suppl.), S139—S165.

Cohen, M. D., March, J. G. & Olsen, J. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, *17*(1), 1—25.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London: UCL Press.

D'Aquino, P., Le Page, C., Bousquet, F., Bah, A. (2003). Using self-designed role-playing games and a multi-agent system to empower a local decision-making process for land use management: The selfCormas experiment in Senegal. *Journal of Artificial Societies and Social Simulation*, *6*(3). Retrieved from http://jasss.soc.surrey.ac.uk/6/3/5.html.

Dean, J. S., Gumerman, G. j., Epstein, J. M., Axtell, R. L., Swedland, A. C., Parker, M. T., & McCarroll, S. (1999). Understanding Anasazi culture change through agent-based modeling. In T. Kohler & G. Gumerman(Eds.), *Dynamics in human and primate societies*. Oxford, UK: Oxford University Press.

Deffuant, G. (2006). Comparing extremism propagation patterns in continuous opinion models. *Journal of Artificial Societies and Social Simulation*, *9*(3). Retrieved from http://jasss.

soc. surrey. ac. uk/9/3/8. html.

Degushi, H. (2004). *Economics as an agent-based complex system*. Tokyo: Springer-Verlag.

Dibble, C. , & Feldman, P. G. (2004). The GeoGraph 3D Computational Laboratory: Network and terrain landscapes for RePast. *Journal of Artificial Societies and Social Simulation*, *7*(1). Retrieved from http://jasss. soc. surrey. ac. uk/7/1/7. html.

Dray, A. , Perez, P. , Jones, N. , Le Page, C. , D'Aquino, P. , White, I. , & Auatabu, T. (2006). The AtollGame experience: From knowledge engineering to computer-assisted role playing game. *Journal of Artificial Societies and Social Simulation*, *9*(1). Retrieved from http://jasss. soc. surrey. ac. uk/9/1/6. html.

Dunham, J. B. (2005). An agent-based spatially explicit epidemiological model in MASON. *Journal of Artificial Societies and Social Simulation*, *9*(1). Retrieved from http://jasss. soc. surrey. ac. uk/9/1/3. html.

Eckel, B. (2005). Thinking in Java. Englewood Cliffs, NJ: Prentice Hall.

Edmonds, B. (2006). The emergence of symbiotic groups resulting from skill-differentiation and tags. *Journal of Artificial Societies and Social Simulation*, *9*(1). Retrieved from http:// jasss. soc. surrey. ac. uk/9/1/10. html.

Edmonds, B. , & Hales, D. (2003). Replication, replication and replication: Some hard lesions from model alignment. *Journal of Artificial Societies and Social Simulation*, *6*(4). Retrieved from http://jasss. soc. surrey. ac. uk/6/4/11. html.

Eiben, A. E. , & Smith, J. E. (2003). *Introduction to evolutionary computing*. New York: Springer.

Elias, N. (1969). *The civilizing process*. Oxford, UK: Blackwell. (Original work published 1939).

Engelbrecht, A. P. (2002). *Computational intelligence: An introduction*. Chichester, UK: Wiley.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41—60.

Epstein, J. M. (2007). *Generative social science: Studies in agent-based computational modeling*. Princeton, NJ: Princeton University Press.

Epstein, J. M. , & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Washington, DC: Brookings Institution Press.

Erev, I. , & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, *88*(4), 848—881.

Etienne, M. (2003). SYLVOPAST: A multiple target role-playing game to assess negotiation process in sylvopastoral management planning. *Journal of Artificial Societies and Social Simulation*, *6*(2). Retrieved from http://jasss. soc. surrey. ac. uk/6/2/5. html.

Etienne, M. , Le Page, C. , & Cohen, M. (2003). A step-by-step approach to building land management scenarios based on multiple viewpoints on multi-agent system simulations. *Journal of Artificial Societies and Social Simulation*, 6(2). Retrieved from http://jasss. soc. surrey. ac. uk/6/2/2. html.

Fagiolo, G. , Windrum, P. , & Moneta, A. (2006). *Empirical validation of agent-based models:*

A critical survey(No. 2006/14). Pisa, Italy: Sant'Anna School of Advanced Studies, Laboratory of Economics and Management.

Farmer, J. D., Patelli, P., & Zovko, L. I. (2005). The predicative power of zero intelligence in financial markets. *Proceedings of the National Academy of Sciences*, *102*(6), 2254—2259. Retrieved from http://www.santafe.edu/~jdf/papers/zero.pdf.

Fielding, j., & Gilbert, N. (2000). *Understanding social statistics*. London: Sage.

Fink, E. C., Gates, S., & Humes, B. D. (1998). *Game theory topics: Incomplete information, repeated games and n-player games*(Sage Series on Quantitative Applications in the Social Sciences, Vol, 122). Thousand Oaks, CA: Sage.

Fioretti, G. (2001). Information structure and behavior of a textile industrial district. *Journal of Artificial Societies and Social Simulation*, *4*(4). Retrieved from http://jasss.soc.surrey.ac.uk/4/4/1.html.

Flache, A., & Macy, M. (2002). Stochastic collusion and the power law of learning. *Journal of Conflict Research*, *46*(5), 629—653.

Forrester, J. W. (1971). *World Dynamics*. Cambridge: MIT Press.

Fortheringhan, A. S., & O'Kelly, M. E. (1989). *Spatial interaction models: Formulations and applications*. Dordrecht: Kluwer.

Friedman-Hill, E. (2003). *Jess in action: Rule-based system in Java*. Greewich, CT: Manning.

Galan, J. M., & Izquierdo, L. (2005). Appearance can be deceiving: Lessons learned re-implementing Axelrod's "evolutionary approach to norms." *Journal of Artificial Societies and Social Simulation*, *8*(3). Retrieved from http://jasss.soc.surrey.ac.uk/8/3/2.html.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. London: Sage.

Gaylord, R. J., & D'Andria, L. (1998). *Simulating society: A mathematica toolkit for modeling socioeconomic behavior*. New York: TELOS/Springer Verlag.

Gilbert, N. (1997). A simulation of the structure of academic science. *Sociological Research Online*, *2*(2). Retrieved from http://www.socresonline.org.uk/socresonline/2/2/3.html.

Gilbert, N. (2002). Varieties of emergence. In D. Sallach(Ed.), *Agent* 2002: *Social agents: Ecology, exchange, and evolution*(pp. 41—56). Chicago: University of Chicago and Argonne National Laboratory.

Gilbert, N. (2006). *A generic model of collectivities*. ABModSim 2006. International Symposium on Agent Based Modeling and Simulation. University of Vienna: European Meeting on Cybernetic Science and System Research.

Gilbert, N., & Abbott, A. (Eds.)(2005). Social Science computation [Special issue]. *American Journal of Sociology*, *110*(4) Chicago: University of Chicago Press.

Gilbert, N., & Bankes, S. (2002). Platforms and methods for agent-based modeling. *Proceedings of the National Academy of Sciences*, *99*(Suppl. 3), 7197—7198.

Gilbert, N., den Besten, M., Bontovics, A., Craenen, B. G. W., Divina, F., Eiben, A. E., et al. (2006). Emerging artificial societies through learning. *Journal of Artificial Societies and Social Simulation*, *9*(2). Retrieved from http://jasss.soc.surrey.ac.uk/9/2/9.html.

Gilbert, N. , Pyka, A. , & Ahrweiler, P. (2001). Innovation networks: A simulation approach. *Journal of Artificial Societies and Social Simulation*, *4*(3). Retrieved from http://jasss.soc.surrey.ac.uk/4/3/8.html.

Gilbert, N. , & Terna, P. (2000). How to build and use agent-based models in social science. *Mind and Society*, *1*(*1*), 57—72.

Gilbert, N. , & Troitzsch, K. G. (2005). *Simulation for the social scientists* (2nd ed.). Milton Keynes, UK: Open University Press.

Gimblett, H, R. (2002). *Integrating geographic information systems and agent-based modeling techniques for simulating social and ecological processes*. London: Oxford University Press.

Goolsby, R. (2006). Combating terrorist networks: An evolutionary approach. *Computational and Mathematical Organization Theory*, *12*, 7—20.

Grimm, V. , & Railsback, S. F. (2005). *Individual-based modeling and ecology*. Princetion, NJ: Princeton University Press.

Gupta, A. , & Kapur, V. (2000). *Microsimulation in government policy and forecasting*. Amsterdam: Elsevier.

Hales, D. (2000). Evolving specialization, altruism, and group-level optimization using tags. In J. S. Sichman, F. Bousquet, & P. Davidsson(Eds.). *Multi-agent-based simulation* Ⅱ (pp. 26—35). Berlin: Springer-Verlag.

Harding, A. (1996). *Microsimulation and public policy*. New York: Elsevier.

Hegslmann, R. , & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3). Retrieved from http://jasss.soc.surrey.ac.uk/5/3/2.html.

Heywood, D. I. , Cornelius, S. , & Carver, S. (2002). *An introduction to geographical information systems* (2nd ed.). New York: Prentice Hall.

Hodkinson, P. (2002). *Goth identity, style and subculture*. Oxford, UK: Berg.

Homer, J. B. , & Hirsch, G. B. (2006). System dynamics modeling for public health: Background and opportunities. *American Journal of Public Health*, *96*(3), 452—458.

Huberman, B. A. , & Glance, N. (1993). Evolutionary games and computer simulations. *Proceeding of the National Academy of Sciences*, *90*, 7716—7718.

Hutchins, E. , & Hazlehurst, B. (1995). How to invent a lexicon: The development of shared symbols in interaction. In N. Gilbert & R. Conte(Eds.), *Artificial Societies*. London: UCL Press.

Izquierdo, S. S. , & Izquierdo, L. R. (2006). *The impact on market efficiency of quality uncertainty without asymmetric information*. Paper presented at the Workshop on Agent-Based Models of Consumer Behavior and Market Dynamics, Guildford, UK.

Jacobsen, C. , Bronson, R. , & Operations Research Society of America. (1985). *Simulation violators*. Arlington, VA: Operations Research Society of America.

Janssen, M. ,& Jager, W. (1999). An integrated approach to simulating behavioral processes: A case study of the lock-in of consumption patterns. *Journal of Artificial Societies and Social Simulation*, *2*(2). Retrieved from http://jasss.soc.surrey.ac.uk/2/2/2.html.

Johnson, P. E. (2002). Agent-based modeling: what I learned from the artificial stock market. *Social Science Computer Review*, *20*, 174—186.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449—1475.

Kaldor, N. (1961). Capital accumulation and economic growth. In F. A. Lutz & D. C. Hague (Eds.), *The theory of capital* (pp. 177—222). London: Macmillan.

Kitts, J. A. Macy, M. W., & Flache, A. (1999). Structural learning: Attraction and conformity in task-oriented groups. *Computational & Mathematical Organization Theory*, *5*(2), 129—145.

Klemm, K., Eguíluz, V. M., Toral, R., & San Miguel, M. (2003). Global culture: A noise-induced transition in finite systems. *Physcial Review E*, *67*(4). Retrieved from http://arxiv.org/abs/cond-mat/0205188.

Koesrindartoto, D., Sun, J., & Tesfatsion, L. S. (2005). *An agent-based computational laboratory for testing the economic reliability of wholesale power market designs*. Paper presented at eh Energy, Environment, and Economics in a New Era conference, San Francisco.

Koza, J. R. (1992). *Genetic programming*. Cambridge: MIT Press.

Koza, J. R. (1994). *Genetic programming 2*. Cambridge: MIT Press.

Laird, J. K., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1—64.

LeBaron, B., Arthur, W. B., & Palmer, R. (1999). Time series properties of an artificial stock market. *Journal of Economic Dynamic and Control*, *23*, 1487—1516.

Liebrand, B. G., Nowak, A., & Hegselmann, R. (Eds.). (1998). *Computer modeling of social process*. London: Sage.

Link, J., & Fröhlich, P. (2003). *Unit testing in Java: How tests drive the code*. San Francisco: Morgan Kaufmann.

Lorenz, J. (2006). Consensus strikes back in the Hegselmann-Krause Model of continuous opinion dynamics under bounded confidence. *Journal of Artificial Societies and Social Simulation*, 9(1). Retrieved from http://jasss.soc.surrey.ac.uk/9/1/8.html.

Luke, S., Cioffi-Revilla, C., Panait, L., Sullivan, K., & Balan, G. (2005). MASON: A Java multi-agent simulation environment. *Simulation: Transactions of the Society for Modeling and Simulation International*, *81*(7), 517—527.

Macy, M., & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*, 143—166.

Malerba, F., Nelson, R., Orsenigo, L., & Winter, S. G. (2001). History-friendly models: An overview of the case of the computer industry. *Journal of Artificial Societies and Social Simulation*, *4*(3). Retrieved from http://jasss.soc.surrey.ac.uk/4/3/6.html.

Marengo, L. (1992). Coordination and organizational learning in the firm. *Journal of Evolutionary Economics*, *2*(3), 313—326.

McKeown, G., & Sheehy, N. (2006). Mass media and polarization processes in the bounded confidence model of opinion dynamics. *Journal of Artificial Societies and Social Simulation*, *9*(1). Retrieved from http://jasss.soc.surrey.ac.uk/9/1/11.html.

Merton, R. K. (1968). *Social theory and social structure*(Enl. ed.) New York: Free Press.

Meyer, M., & Hufschlag, K. (2006). A generic approach to an object-oriented learning classifier system library. *Journal of Artificial Societies and Social Simulation*, *9*(3). Retrieved from http://jasss.soc.surrey.ac.uk/9/3/2.html.

Miles, R. (2006). *Learning UML 2.0*. Sebastopol, CA: O'Reilly.

Miiton, L., Sutherland, H., & Weeks, M. J. (2000). *Microsimulation modeling for policy analysis: Challenges and innovations*. Cambridge, UK: Cambridge University Press.

Mookherjee, D., & Sopher, B. (1994). Learning behavior in an experimental matching pennies game. *Games and Economic Behavior*, *7*, 62—91.

Mookherjee, D., & Sopher, B. (1997). Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, 19—132.

Moss, S. (2002). Policy analysis from first principles. *Proceedings of the National Academy of Sciences*, *99*(Suppl. 3), 7267—7274.

Mulkay, M. J., & Turner, B. S. (1971). Over-production of personnel and innovation in three social settings. *Sociology*, *5*(1), 47—61.

Niemeyer, P., & Knudsen. J. (2005). *Learning Java*(3rd ed.). Sebastopol, CA: O'Reilly.

Nilsson, N. (1998). *Artificial intelligence: A new synthesis*. San Francisco: Morgan Kaufmann.

North, M. J. (2001). Multi-agent social and organizational modeling of electric power and natural gas markets. *Computational and Mathematical Organization Theory*, *7*. 331—337.

North, M. J., Collier, N. T., & Vos, J. R. (2006). Experience creating three implementations of the repast agent modeling toolkit. *ACM Transactions on Modeling and Computer Simulation*. *16*(1), 1—25.

Orcutt, G., Merz, J., & Quinke, H. (1986). *Microanalytic simulation models to support social and financial policy*. Amsterdam: North-Holland.

O'Suillvan, D., & Macgill, J. (2005). *Modeling urban residential neighborhood dynamics*. Paper presented at the workshop on Modeling Urban Social Dynamics, University of Surrey, Guildford, UK.

Pajares, J., Hernández-Iglesias, C., & López-Paredes, A. (2004). Modeling learning and R & D in innovative environments: A cognitive multi-agent approach. *Journal of Artificial Societies and Social Simulation*, *7*(2). Retrieved from http://jasss.soc.surrey.ac.uk/7/2/7.html.

Pancs, R., & Vriend, N. J. (2004). *Schelling's spatial proximity model of segregation revisited*. International Workshop on "Agent-based Models for Economic Policy Design"(ACEPOL05) Retrieved January 20, 2007, from http://ww.wiwi,uni-bielefeld.de/~dawid/acepol/Downloads/Paper/Vriend.pdf.

Papert, S. (1980). *Mindstroms: Children, computers, and powerful ideas*. New York: Basic Books.

Parunak, H. V. D., Savit, R., & Riolo, R. L. (1998, July), Agent-based modeling vs. equation-based modeling: A case study and users' guide. In J. S. Sichman, R. Conte, & N. Gilbert (Eds.), *Proceedings of Multi-Agent Systems and Agent-Based Simulation*(pp. 10-25). Paris: Springer.

Perfors, A. (2002). Simulated evolution of language: A review of the field. *Journal of Artificial Societies and Social Simulation*, *5*(2). Retrieved from http://jasss.soc.surrey.ac.uk/5/2/4.html.

Phillips, A. W. (1950). Mechanical models in economic dynamics. *Economica*, *17*(67), 283—305.

Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). Network dynamics and field evolution: The growth of inter-organizational collaboration in the life sciences. *American Journal of Sociology*, *110*(4), 1132—1205.

Pujol, J. M., Flache, A., Delgado, J. & Sangüesa, R. (2005). How can social network ever become complex? Modeling the emergence of complex networks from local social exchanges. *Journal of Artificial Societies and Social Simulation*, *8*(4). Retrieved from http://jasss.soc.surrey.ac.uk/8/4/2.html.

Railsback, S. F., Lytinen, S. L., & Jackson, S. K. (2006). Agent-based simulation platforms: Review and development recommendations. *Simulation: Transactions of the Society for Modeling and Simulation International*, *82*(9), 609—623. Available: http://www.humboldt.edu/~ecomodel/documents/ABMPlatformReview.pdf.

Ramanath, A. M., & Gilbert, N. (2004). Techniques for the construction and evaluation of parrticipatory simulations. *Journal of Artificial Societies and Social Simulation*, *7*(4). Retrieved from http://jasss.soc.surrey.ac.uk/7/4/1.html.

Raney, B., Cetin, N., Vollmy, A., & Nagel, K. (2002). *Large scale multi-agent transportation simulations*. Paper presented at the Proceedings of the Annual Congress of the European Regional Science Association(ERSA), Dortmund, Germany. Available: http://www.inf.ethz.ch/nagel/papers.

Redmond, G., Sutherland, H., & Wilson, M. (1998). *The arithmetic of tax and social security reform: A user's guide to microsimulation methods and analysis*. Cambridge, UK: Cambridge University Press.

Reynolds, C. W. (1987). Rocks, herds, and schools: A distributed behavioral model. *Computer Graphics*, *21*(4), 25—34.

Richiardi, M., Leombruni, R., Saam, N., & Sonnessa, M. (2006). A common protocol for agent-based social simulation. *Journal of Artificial Societies and Social Simulation*, *9*(1). Retrieved from http://jasss.soc.surrey.ac.uk/9/1/15.html.

Riolo, R. L., Cohen, M. D., & Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, *411*, 441—443.

Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, *8*, 164—212.

Rouchier, J. (2003). Re-implementation of a multi-agent model aimed at sustaining experimenntal economic research: The case of simulations with emerging speculation. *Journal of Artiificial Societies and Social Simulation*, *6*(4). Retrieved from http://jasss.soc.surrey.ac.uk/6/4/7.html.

Sakoda, J. M. (1971). The checkerboard model of social interaction. *Journal of Mathematical Sociology*, *1*(1), 119-131.

Sallans, B., Pfister, A., Karatzoglou, A., & Dorffner, G. (2003). Simulation and validation of an integrated markets model. *Journal of Artificial Societies and Social Simulation*, *6*(4).

Retrieved from http://jasss. soc. surrey. ac. uk/6/4/2. html.

Sawyer, R. K. (2004). Social explanation and computational simulation. *Philosophical Explorations*, 7(3), 219—231.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1, 143—186.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. New York: Norton.

Schweitzer, F. (2003). *Brownian agents and active particles*. New York: Springer-Verlag.

Scott, J. (2000). *Social network analysis*(2nd ed.). London: Sage.

Simmel, G. (1907). Fashion. *International Quarterly*, 10, 130—155.

Simon, H. A. (1957). A behavioral model of rational choice. In H. A. Simon(Ed.), *Models of man*. New York: Wiley.

Slanina, F. (2000). *Social organization in the minority game model*. Retrieved from http://arxiv. org/abslcond-mat/0006098.

Squazzoni, F., & Boero. R. (2002). At the edge of variety and coordination: An agent-based computational model of industrial districts. *Journal of Artificial Societies and Social Simuulation*, 5(1). Retrieved from http://jasss. soc. surrey. ac. uk/5/l/l. html.

Stauffer, D., Sousa, A., & Schulze, C. (2004). Discretized opinion dynamics of the Deffaunt model on scale-free networks. *Journal of Artificial Societies and Social Simulation*, 7(3). Retrieved from http://jasss. soc. surrey. ac. uk/7 /3/7. html.

Steels, L. (2004). The evolution of communication systems by adaptive agents. In E. Alonso, D. Kudenko, & D. Kazakov(Eds.), *Adaptive agents and multi-agent systems* (Vol. 2636, pp. 125—140). Berlin: Springer Verlag.

Sterman, J. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Boston: Irwin McGraw-Hill.

Strader, T. J., Lin, F., & Shaw, M. J. (1998). Simulation of order fulfillment in divergent assembly supply chains. *Journal of Artificial Societies and Social Simulation*, 1 (2). Retrieved from http://jasss. soc. surrey. ac. uk/1/2/5. html.

Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In R. Sun(Ed.) *Cognition and multi-agent interaction: From cognitive modeling to social simulation*. New York: Cambridge University Press.

Sun, R., & Naveh, I. (2004). Simulating organizational decision-making using a cognitively realistic agent model. *Journal of Artificial Societies and Social Simulation*, 7(3). Retrieved from http://jasss. soc. surrey . ac. uk/7 /3/5. html.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.

Taatgen, N., Lebiere, C., & Anderson, J. (2006). Modeling paradigms in ACT-R. In R. Sun (Ed.), *Cognition and multi-agent interaction: From cognitive modeling to social simulaation*. Cambridge, UK: Cambridge University Press.

Tesfatsion, L., & Judd, K. (2006). *Handbook of computatiollal economics*(Vol. 2). Amsterdam: North-Holland.

Thorngate, W. (2000). Teaching social simulation with Matlab. *Journal of Artificial Societies and Social Simulation*, *3*(1). Retrieved from http://www. soc. surrey. ac. uk/JASSS/3/1/ forum/1. html.

Thornton, S. (1995). *Club cultures: Music, media and subcultural capital*. Cambridge, MA: Polity.

Tobias, R., & Hofmann, C. (2004). Evaluation of free Java-libraries for social-scientific agent based simulation. *Journal of Artificial Societies and Social Simulation*, *7*(1). Retrieved from http://jasss. soc. surrey. ac. uk/7/l/6. html.

Troitzsch, K. G. (2004). *Validating simulation models*. 18th European Simulation Multiconference, SCS Europe.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167*(1—2), 206—242.

Vogt, P., & Coumans, H. (2003). Investigating social interaction strategies for bootstrapping lexicon development. *Journal of Artificial Societies and Social Simulation*, *6*(1). Retrieved from http://jasss. soc. surrey. ac. uk/6/1/4. html.

Waterman, D. A., & Hayes-Roth, F. (1978). *Pattern-directed inference systems*. Orlando, FL: Academic Press.

Watts, D. (1999). Network dynamics and the small world phenomenon. *American Journal of Sociology*, *105*(2), 493—527.

Watts, D., & Strogatz, S. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*, 440—442.

Widdicombe, S., & Wooffitt, R. C. (1990). "Being" versus "doing" punk(etc): On achieving authenticity as a member. *Journal of Language and Social Psychology*, *9*, 257—277.

Wilensky, U. (1998). NetLogo segregation model. Evanston, IL: Northwestern University, Center for Connected Learning and Computer-Based Modeling. Retrieved from http://ccl. northwestern. edu/netlogo/models/Segregation.

Wilensky, U. (1999). *NetLogo*. Evanston, IL: Northwestern University, Center for Connected Learning and Computet-Based Modeling. Retrieved from http://ccl. northwestern. edu/netlogo.

Wilenskv, U. (2005). *NetLogo wolf sheep predation (system dynamics) model*. Evanston, IL: Nonhwestern University, Center for Connected Learning and Computer-Based Modeling. Retrieved from http://ccl. northwestern. edu/netlogo/models/WolfSheepPredaticn (System Dynamics)

Wooldridge, M., & Iennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review*, *10*(2), 115—152.

Wray, R. E., & lones, R. M. (2006). Considering Soar as an agent architecture. In R. Sun(Ed.), *Cognition and multi-agent interaction: From cognitive modeling to social simulation*. Cambridge, UK: Cambridge University Press.

Ye, M., & Carley, K. M. (1995). Radar-Soar: Towards an artificial organization composed of intelligent agents. *Journal of Mathematical Sociology*, *20*(2—3), 219—246.

Zhang, I. (2004). A dynamic model of residential segregation. *Journal of Mathematical Sociology*, *28*(3), 147—170.

译名对照表

abstract model	抽象模型
agent-based models	基于行动者模型
agents	行动者
analogical model	类比模型
artificial neural networks(ANNs)	模拟神经网络
automomy	自主性
back-propagation algorithm	反向传播算法
Boid models	Boid 模型
boundedly rational	有限理性
calibrating time	校准时间
companion modeling	伴随建模
connection weight	连接权重
corner testing	拐角测试
debugging	除错
environment	环境
equation-based model	基于公式模型
event-Driven Simulation	事件驱动模拟
evolutionary computation	进化计算
experiment	试验
Facsimile Models	传真模型
Framework	框架
game theory	博弈理论
genetic algorithm	遗传算法
genetic programming	遗传编程
geographical information system(GIS)	地理信息系统
global variables	全局变量
ideal-type model	理想型模型
innovation models	创新模型
learning algorithm	学习算法
learning classifier systems	学习分类器系统
library	程序库
local-radius	局部一半径
machine learning	机器学习
matrix of conditional probabiliteis	条件概率矩阵
microsimulation model	微观模拟模型
middle range model	中层模型

model of innovation networks	创新网络模型
modeling environments	建模环境
modularity	模块性
multiagent system	多行动者系统
NetLogo segregation model	NetLogo 隔离模型
opinion dynamic model	民意动态模型
paricipative modeling	参与建模
power law	幂次法则
preferential attachment	偏好连接
proactivity	主动性
production rule system	生产准则系统
pseudo-code	伪编码
random asynchronous execution	随机不同步执行
random network	随机网络
reactivity	反应性
reinforcement learning	强化学习
role-playing game	角色扮演游戏
rule of thumb	经验法则
scale model	比例模型
sensitivity analysis	灵敏度分析
sequential asynchronous execution	有序不同步执行
simulated synchronous execution	模拟的同步执行
"small world" network	"小世界"网络
social network analysis	社会网络分析
spatially explicit	空间直观
statistical signatures	统计信号
stylized facts	程式化事实
synchronicity	共时性
system dynamics	系统动力学
target of model	模型原型
the minority game	少数人游戏
toolkits	工具箱
unified modeling languages(UML)	统一建模语言
validation	有效性验证
verification	检验
visible-radius	可见一半径
working memory	工作记忆